Volume 38 Number 2 June 2014

ISSN 0350-5596

Informatica

An International Journal of Computing and Informatics

1977

Editorial Boards

Informatica is a journal primarily covering intelligent systems in the European computer science, informatics and cognitive community; scientific and educational as well as technical, commercial and industrial. Its basic aim is to enhance communications between different European structures on the basis of equal rights and international refereeing. It publishes scientific papers accepted by at least two referees outside the author's country. In addition, it contains information about conferences, opinions, critical examinations of existing publications and news. Finally, major practical achievements and innovations in the computer and information industry are presented through commercial publications as well as through independent evaluations.

Editing and refereeing are distributed. Each editor from the Editorial Board can conduct the refereeing process by appointing two new referees or referees from the Board of Referees or Editorial Board. Referees should not be from the author's country. If new referees are appointed, their names will appear in the list of referees. Each paper bears the name of the editor who appointed the referees. Each editor can propose new members for the Editorial Board or referees. Editors and referees inactive for a longer period can be automatically replaced. Changes in the Editorial Board are confirmed by the Executive Editors.

The coordination necessary is made through the Executive Editors who examine the reviews, sort the accepted articles and maintain appropriate international distribution. The Executive Board is appointed by the Society Informatika. Informatica is partially supported by the Slovenian Ministry of Higher Education, Science and Technology.

Each author is guaranteed to receive the reviews of his article. When accepted, publication in Informatica is guaranteed in less than one year after the Executive Editors receive the corrected version of the article.

On the Inverse Problem for Generalized One-Sided Concept Lattices

Jozef Pócs
Palacký University Olomouc, Department of Algebra and Geometry, Olomouc, Czech Republic
and Mathematical Institute, Slovak Academy of Sciences, Košice, Slovakia
E-mail: pocs@saske.sk

Jana Pócsová
Technical University of Košice, BERG Faculty
Institute of Control and Informatization of Production Processes, Košice, Slovakia
E-mail: jana.pocsova@tuke.sk

Keywords: formal context, generalized one-sided concept lattice, Galois connection, closure system

Received: August 3, 2013

Generalized one-sided concept lattices represent a generalization of the classical concept lattices convenient for analysis of object-attribute models with different types of attributes. Formally, to each object-attribute model (represented by the notion of formal context) there is assigned a pair of concept-forming operators. Fixed points of these operators form a hierarchical structure consisting of extent-intent pairs. From the algebraic point of view this structure forms a complete lattice, called the generalized one-sided concept lattice. In this paper we deal with the inverse problem for generalized one-sided concept lattices. For a given generalized one-sided concept lattice we describe an algorithm for finding the corresponding formal context.

Povzetek: Predstavljen je algoritem za preslikavo enostranske mreže konceptov v pripadajoči formalni koncept.

1 Introduction

In mathematics, physics, computer science or engineering there are pairs of problems which are inverses of one another. As examples from mathematics we can mention the multiplication of integers and as the corresponding inverse problem the factorization of a given integer, differentiation and integration of real valued functions or Fourier transform and inverse Fourier transform. From physics we can mention scattering problem, which is to determine how radiation or particles are scattered based on the characteristics of some object (scatterer) and inverse scattering problem of determining characteristics of an object based on data of how it scatters incoming radiation or particles.

At first glance, the meaning of the term 'inverse problem' seems obvious. It is the problem which is associated to some other (direct) problem, one that presumably preceded the inverse problem and which has been studied extensively for some time and is better known. Our inverse problem concerns determination of characteristics of object-attribute models in fuzzy modification of Formal Concept Analysis (FCA), so called generalized one-sided concept lattices.

The theory of concept lattice, also called Formal Concept Analysis is a theory of data analysis for identification of conceptual structures among data sets. As an effective tool for data analysis, Formal Concept Analysis has been extensively applied to a variety of fields such as data mining, decision making, information retrieval, machine learning and knowledge discovery. The main notion of this theory is the notion of a formal context, represented by a binary relation between the set of objects and the set of attributes, specifying which objects have what attributes. From a formal context, one can construct object-attribute pairs known as the formal concept. The family of all formal concepts forms an algebraic structure called the concept lattice, which reflects the relationship of generalization and specialization among particular concepts. The reader can find an extensive account of the mathematical foundations of FCA in [7].

In many real applications, however, the relationship may be many-valued (fuzzy). Therefore, some attempts have recently been devoted to introduce fuzzy concept lattice with properties similar to the classical ones. We mention approaches [2, 3] based on residuated lattices or multi-adjoint concept lattices [11]. A very important class of fuzzy concept lattices is formed by the one-sided concept lattices, where usually objects are considered as crisp subsets and attributes obtain fuzzy values, cf. [9] or [10]. In this case interpretation of object clusters is straightforward as in classical FCA. Consequently, all known applications developed for classical concept lattices can be used in the theory of one-sided concept lattices. Recently there was a generalization of all one-sided approaches (the so-called generalized one-sided concept lattices, see [6] for more details), which allows one to consider different types of structure for truth degr ees (represented by complete lattices). From this point of view it is applicable to a very wide spectrum of the real object-attribute models where methods of the classical FCA are appropriate, cf. [1, 5, 4, 8, 14, 15].

As we have already mentioned, our aim is to deal with the inverse problem for generalized one-sided concept lattices. The paper is organized as follows: in the next section we give a brief overview of the notions concerning generalized one-sided concept lattices. We recall some algebraic notions like Galois connections, complete lattices or closure systems. Our main result, i.e., an algorithm for the inverse problem is presented in Section 3. In particular we will deal with the decision problem, i.e., whether a given collection of pairs forms a generalized one-sided concept lattice, and consequently we describe a method for determining the formal context (object-attribute model) corresponding to a given generalized one-sided concept lattice.

2 Generalized one-sided concept lattices

In this section we describe a fuzzy generalization of classical concept lattices, the so-called generalized one-sided concept lattices, cf. [6] and [13].

The main idea of fuzzifications of classical FCA is the usage of graded truth. The structure L of truth degrees forms a so-called complete lattice, i.e., it is partially ordered, contains the smallest and the greatest element (representing the values **false** and **true**, respectively), moreover, for any subset $H \subseteq L$ there exists $\bigvee H$ (the least upper bound or supremum) and $\bigwedge H$ (the greatest lower bound or infimum). In classical logic, each proposition is either true or false, hence classical logic is bivalent. It is common to represent the classical logic truth value structure as a two-element chain, i.e., the two-element set $\{0, 1\}$ with $0 < 1$. In this case the value 0 represents false and 1 represents true. In fuzzy logic, to each proposition there is assigned a truth degree from some richer scale L of truth degrees. If to the propositions Φ and Ψ are assigned truth degrees $\parallel \Phi \parallel = a$ and $\parallel \Psi \parallel = b$, then $a \leq b$ means that Φ is considered less true than Ψ. In object-attribute models the typical propositions are of the form "object has attribute in degree a". The well-known examples of truth structures used in various modifications of fuzzy logic are: the real unit interval $[0, 1]$, Boolean algebras, MV algebras, or, more generally, residuated lattices.

The set of all L-fuzzy sets over some universe U is defined as the set of all functions

$$f \colon U \to L,$$

denoted by symbol L^U. In order to define generalized one-sided concept lattices we will use the notion of direct product. If L_i for $i \in I$ is a family of lattices the *direct product* $\prod_{i\in I} L_i$ is defined as the set of all functions

$$f : I \to \bigcup_{i\in I} L_i$$

such that $f(i) \in L_i$ for all $i \in I$ with the "componentwise" order, i.e, $f \leq g$ if $f(i) \leq g(i)$ for all $i \in I$. If $L_i = L$ for all $i \in I$ we get the direct power L^I. In this case the direct power L^I represents the structure of L-fuzzy sets, hence the direct product of lattices can be seen as a generalization of the notion of L-fuzzy sets. The direct product of lattices forms a complete lattice if and only if all members of the family are complete lattices. Straightforward computations show that the lattice operations in the direct product $\prod_{i\in I} L_i$ of complete lattices are calculated componentwise, i.e., for any subset $\{f_j : j \in J\} \subseteq \prod_{i\in I} L_i$ we obtain

$$(\bigvee_{j\in J} f_j)(i) = \bigvee_{j\in J} f_j(i),$$

$$(\bigwedge_{j\in J} f_j)(i) = \bigwedge_{j\in J} f_j(i),$$

where these equalities hold for each index $i \in I$.

In order to introduce the notion of generalized one-sided concept lattices as a generalization of FCA we will assume only one minimal condition, i.e., that the structures of truth degrees form complete lattices.

In the mathematical theory of fuzzy concept lattices, the main role is played by special pairs of mappings between complete lattices, commonly known as Galois connections. Hence, we provide necessary details regarding Galois connections and related topics.

Let $(L, \leq)$ and $(M, \leq)$ be complete lattices and let $\varphi \colon L \to M$ and $\psi \colon M \to L$ be maps between these lattices. Such a pair (φ, ψ) of mappings is called a *Galois connection* if the following condition is fulfilled:

$$p \leq \psi(q) \quad \text{if and only if} \quad \varphi(p) \geq q.$$

Galois connections between complete lattices are closely related to the notion of closure operator and closure system. Let L be a complete lattice. By a *closure operator* in L we understand a mapping $c \colon L \to L$ satisfying:

(a) $x \leq c(x)$ for all $x \in L$,

(b) $c(x_1) \leq c(x_2)$ for $x_1 \leq x_2$,

(c) $c(c(x)) = c(x)$ for all $x \in L$ (i.e., c is idempotent).

A subset X of the complete lattice L is called a *closure system* in L if X is closed under arbitrary meets. We note that this condition guarantees that $(X, \leq)$ is a complete lattice, in which the infima are the same as in L, but the suprema in X may not coincide with those from L. For a closure operator c in L, the set $\mathsf{FP}(c)$ of all fixed points of c (i.e., $\mathsf{FP}(c) = \{x \in L : c(x) = x\}$) is a closure system in L. Conversely, for a closure system X in L, the mapping $\mathsf{C}_X : L \to L$ defined by $\mathsf{C}_X(x) = \bigwedge\{u \in X : x \leq u\}$ is a closure operator in L. Moreover these correspondences are inverses of each other, i.e., $\mathsf{FP}(\mathsf{C}_X) = X$ for each closure system X in L and $\mathsf{C}_{\mathsf{FP}(c)} = c$ for each closure operator c in L.

Next we describe the mathematical framework for the generalized one-sided concept lattices. We start with the definition of formal context, from which there is defined a pair of mappings forming a Galois connection.

A 4-tuple $\big(B, A, \mathsf{L}, R\big)$ is said to be a *generalized one-sided formal context* if B is a non-empty set of objects, A is a non-empty set of attributes, $\mathsf{L}\colon A \to \mathsf{CL}$ is a mapping from the set of attributes to the class of all complete lattices. In this case $\mathsf{L}(a)$ represents a particular structure of truth value degrees for each attribute $a \in A$. Finally, $R\colon B \times A \to \bigcup_{a\in A} \mathsf{L}(a)$ with $R(b, a) \in \mathsf{L}(a)$ is an incidence relation, which represents a degree from the structure $\mathsf{L}(a)$ in which an element $b \in B$ has a given attribute $a \in A$.

The power set (set of all subsets) of a set B will be denoted by $\mathbf{P}(B)$. Let $\big(B, A, \mathsf{L}, R\big)$ be a generalized one-sided formal context. Then there is defined a pair of mappings ${}^{\perp}\colon \mathbf{P}(B) \to \prod_{a\in A} \mathsf{L}(a)$ and ${}^{\top}\colon \prod_{a\in A} \mathsf{L}(a) \to \mathbf{P}(B)$ as follows:

$$X^{\perp}(a) = \bigwedge_{b\in X} R(b, a), \tag{1}$$

$$g^{\top} = \{b \in B : \forall a \in A,\ g(a) \leq R(b, a)\}. \tag{2}$$

The pair $({}^{\perp}, {}^{\top})$ forms a Galois connection between $\mathbf{P}(B)$ and L^A. The composition of mappings ${}^{\perp}$ and ${}^{\top}$ forms a closure operator in $\mathbf{P}(B)$ and similarly the composition of ${}^{\top}$ and ${}^{\perp}$ forms a closure operator in $\prod_{a\in A} \mathsf{L}(a)$. Hence, subsets of the form $X^{\perp\top}$ for any $X \subseteq B$ are closed subsets with respect to the closure operator defined above. As it is known, the closed subsets of any closure operator form a complete lattice with respect to the inherited partial order from the underlying complete lattice structure (in this case $\mathbf{P}(B)$). This fact stands behind the formal definition and characterization of concept lattices.

For a given generalized one-sided formal context $\big(B, A, \mathsf{L}, R\big)$ the symbol $\mathfrak{C}\big(B, A, \mathsf{L}, R\big)$ will denote the set of all pairs (X, g) with $X \subseteq B$, $g \in \prod_{a\in A} \mathsf{L}(a)$, satisfying

$$X^{\perp} = g \quad \text{and} \quad X = g^{\top}.$$

In this case, the set X is usually referred to as the *extent* and g as the *intent* of the concept (X, g). Further we define a partial order on $\mathfrak{C}\big(B, A, \mathsf{L}, R\big)$ as follows:

$$(X_1, g_1) \leq (X_2, g_2) \ \text{ iff } \ X_1 \subseteq X_2 \text{ iff } g_1 \geq g_2.$$

Let $\big(B, A, \mathsf{L}, R\big)$ be a generalized one-sided formal context. The set $\mathfrak{C}\big(B, A, \mathsf{L}, R\big)$ with the partial order defined above forms a complete lattice, where

$$\bigwedge_{i\in I} (X_i, g_i) = \Big(\bigcap_{i\in I} X_i, \big(\bigvee_{i\in I} g_i\big)^{\top\perp}\Big)$$

$$\bigvee_{i\in I} (X_i, g_i) = \Big(\big(\bigcup_{i\in I} X_i\big)^{\perp\top}, \bigwedge_{i\in I} g_i\Big)$$

for each family $(X_i, g_i)_{i\in I}$ of elements from $\mathfrak{C}\big(B, A, \mathsf{L}, R\big)$.

The lattice $\mathfrak{C}\big(B, A, \mathsf{L}, R\big)$ is called the *generalized one-sided concept lattice*.

R	a_1	a_2	a_3	a_4
b_1	2	0.25	0	1
b_2	3	0.50	1	0
b_3	1	0.35	0	0
b_4	1	0.25	0	1
b_5	2	0.70	1	0

Table 1: Incidence relation R.

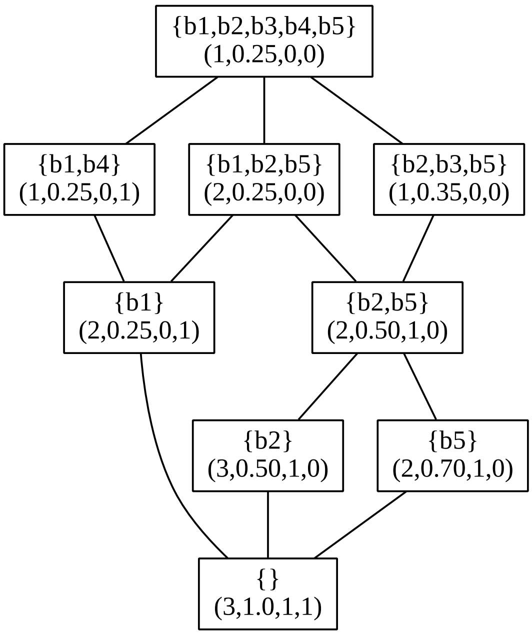


Figure 1: Generalized one-sided concept lattice corresponding to (B, A, L, R).

We provide a small example of a generalized one-sided formal context and the corresponding generalized one-sided concept lattice. Consider the five-element set of objects $B = \{b_1, b_2, b_3, b_4, b_5\}$, and the four-element set of attributes $A = \{a_1, a_2, a_3, a_4\}$ with $\mathsf{L}(a_1) = \mathbf{4}$, $\mathsf{L}(a_2) = [0, 1]$ and $\mathsf{L}(a_3) = \mathsf{L}(a_4) = \mathbf{2}$. In this case $\mathbf{4}$ denotes the four-element chain $0 < 1 < 2 < 3$ and $\mathbf{2}$ denotes the two-element chain $0 < 1$. Finally, the incidence relation R is given in Table 1. Obviously the triple (B, A, L, R) forms a generalized one-sided formal context. The corresponding generalized one-sided concept lattice is depicted in Figure 1.

3 The inverse problem for generalized one-sided concept lattices

After introducing the necessary theoretical background for the direct problem (creation of a generalized one-sided concept lattices from a given formal context), we can provide the precise definition of our inverse problem. Let $B \neq \emptyset$ be a set of objects, $A \neq \emptyset$ be a set of attributes, $\mathsf{L}(a)$ be a system of complete lattices (truth structures under consideration) and $\mathfrak{C}$ be a set consisting of some pairs (X, g) where $X \subseteq B$ and $g \in \prod_{a \in A} \mathsf{L}(a)$. Decide whether there exists (in affirmative case also find) an incidence relation $R \colon B \times A \to \bigcup_{a \in A} \mathsf{L}(a)$ such that $\mathfrak{C} = \mathfrak{C}(B, A, \mathsf{L}, R)$.

In order to decide the inverse problem for generalized one-sided concept lattices we will use the following well-known characterization of Galois connections involving dual isomorphism of closure systems, cf. [12]:

Let L, M be complete lattices. Any Galois connection between L and M is fully determined by dually isomorphic closure systems in L and M.

In order to provide more details, suppose that X_1 and X_2 are closure systems in L, M respectively, and $f : X_1 \to X_2$ is a dual isomorphism between complete lattices $(X_1, \leq)$ and $(X_2, \leq)$. Then a pair $(c_{X_1} \circ f, c_{X_2} \circ f^{-1})$, where c_{X_1}, c_{X_2} are the closure operators corresponding to X_1 and to X_2, forms a Galois connection between L and M.

Given $\mathfrak{C}$ as input, we want to decide if there exists a Galois connection (φ, ψ) between $\mathbf{P}(B)$ and $\prod_{a \in A} \mathsf{L}(a)$ such that

$$\mathfrak{C} = \big\{(X, g) : X = \psi(g), g = \varphi(X)\big\}.$$

If this condition is satisfied, then one can find a corresponding formal context such that $\mathfrak{C} = \mathfrak{C}(B, A, \mathsf{L}, R)$, hence this condition is equivalent to our decision problem. In order to solve this issue, we will use the previous result concerning Galois connections, i.e., we verify that the projections of $\mathfrak{C}$ form dually isomorphic closure systems.

The first step is to decide whether the sets

$$C_1 = \{X \subseteq B : (\exists g)(X, g) \in \mathfrak{C}\}$$

and

$$C_2 = \{g \in \prod_{a \in A} \mathsf{L}(a) : (\exists X)(X, g) \in \mathfrak{C}\}$$

form closure systems in $\mathbf{P}(B)$ and $\prod_{a \in A} \mathsf{L}(a)$, respectively.

Hence we must check if C_1 is closed under arbitrary intersections and C_2 is closed under arbitrary meets.

The second step is to decide whether C_1 and C_2 form dually isomorphic closure systems. We recall that a surjective mapping $f \colon C_1 \to C_2$ is a dual isomorphism if for all $X_1, X_2 \in C_1$ it is true that

$$X_1 \subseteq X_2 \quad \text{iff} \quad f(X_1) \geq f(X_2). \tag{3}$$

This condition will be satisfied if for all $X_1, X_2 \in C_1$ it holds that $X_1 \subseteq X_2$ implies $g_1 \geq g_2$ where g_1 and g_2 are such that $(X_1, g_1) \in \mathfrak{C}$ and $(X_2, g_2) \in \mathfrak{C}$. Similarly, it must hold for all $g_1, g_2 \in C_2$ that $g_1 \geq g_2$ implies $X_1 \subseteq X_2$ where X_1 and X_2 are such that $(X_1, g_1) \in \mathfrak{C}$ and $(X_2, g_2) \in \mathfrak{C}$.

Now we can summarize the whole procedure in the following algorithm (Algorithm 1).

Algorithm 1 for deciding existence of the incidence relation R

```
Input: a set of pairs $\mathfrak{C}$
Output: answer YES or NO
 1: $C_1 \leftarrow \{X : (X, g) \in \mathfrak{C}\}$            ▷ Set of first components
 2: $C_2 \leftarrow \{g : (X, g) \in \mathfrak{C}\}$            ▷ Set of second components
 3: for all $X_1, X_2 \in C_1$ do
 4:     if $X_1 \cap X_2 \notin C_1$ then            ▷ $C_1$ is not a closure system
 5:         return NO
 6:     end if
 7: end for
 8: for all $g_1, g_2 \in C_2$ do
 9:     if $g_1 \wedge g_2 \notin C_2$ then            ▷ $C_2$ is not a closure system
10:         return NO
11:     end if
12: end for
13: for all $X_1, X_2 \in C_1$ such that $X_1 \subseteq X_2$ do
14:     $g_1 \leftarrow g$ where $(X_1, g) \in \mathfrak{C}$, $g_2 \leftarrow g$ where $(X_2, g) \in \mathfrak{C}$
15:     if $g_1 \ngeq g_2$ then            ▷ $C_1$ and $C_2$ are not dually isomorphic
16:         return NO
17:     end if
18: end for
19: for all $g_1, g_2 \in C_2$ such that $g_1 \geq g_2$ do
20:     $X_1 \leftarrow X$ where $(X, g_1) \in \mathfrak{C}$, $X_2 \leftarrow X$ where $(X, g_2) \in \mathfrak{C}$
21:     if $X_1 \nsubseteq X_2$ then            ▷ $C_1$ and $C_2$ are not dually isomorphic
22:         return NO
23:     end if
24: end for
25: return YES            ▷ $C_1$ and $C_2$ are dually isomorphic
```

The correctness of this algorithm can be proved using the above-mentioned relationship between Galois connections and dually isomorphic closure systems.

In **for all** loop (line 3 - 7) it is decided whether C_1 forms a closure system in $\mathbf{P}(B)$. Similarly, **for all** loop (line 8 - 12) decides whether C_2 forms a closure system in the direct product of lattices $\prod_{a \in A} \mathsf{L}(a)$. If C_1 and C_2 form closure systems, then the next step is to decide whether the correspondence $f \colon X \mapsto g$, $(X, g) \in \mathfrak{C}$ is a dual isomorphism between C_1 and C_2. This is verified in the two **for all** loops (line 13 - 18, line 19 - 24). Let us note that the condition (3) guarantees that the correspondence f is injective. If $f(X_1) = f(X_2)$ then $f(X_1) \geq f(X_2)$ and $f(X_2) \geq f(X_1)$ which yields $X_1 \subseteq X_2$ and $X_2 \subseteq X_1$. Since the inclusion relation is antisymmetric, we obtain $X_1 = X_2$. Moreover, we deal with finite structures only, hence f is surjective too, and consequently it is a bijection.

The algorithm returns the affirmative answer if and only if C_1 and C_2 are dually isomorphic closure systems. In this case, there is a Galois connection between $\mathbf{P}(B)$ and $\prod_{a \in A} \mathsf{L}(a)$ corresponding to the input set $\mathfrak{C}$.

Let $\mathfrak{C}$ be an input set and $n = |\mathfrak{C}|$ denote the number of all pairs in $\mathfrak{C}$. Obviously $|C_1| \leq n$ and $|C_2| \leq n$. Since there are $\binom{n}{2} = \frac{n\cdot(n-1)}{2}$ different two-element subsets of an n-element set, we obtain that two **for all** loops (line 3 - 7, line 8 - 12) have no more than $c \cdot \frac{n(n-1)}{2} \in O(n^2)$ repetitions. Here we assume that the verification whether $X_1 \cap X_2 \in C_1$ and $g_1 \wedge g_2 \in C_2$ can be done in constant time $c \in \mathsf{R}$. Similarly, the time complexity of other two **for all** loops is $O(n^2)$, hence Algorithm 1 is in $O(n^2)$ time complexity class according to the size of the input set n.

In what follows we will deal with the second problem, hence suppose that the decision problem is answered affirmatively. We describe a procedure for finding the incidence relation corresponding to $\mathfrak{C}$. For this purpose we recall the following assertion concerning Galois connections between power sets and direct products of complete lattices, cf. [6].

Let (ϕ, ψ) be a Galois connection between $\mathbf{P}(B)$ and $\prod_{a\in A} \mathsf{L}(a)$. Then there exists a generalized one-sided formal context $\big(B, A, \mathsf{L}, R\big)$ such that $\phi(X) = X^{\perp}$ for all $X \subseteq B$ and $\psi(g) = g^{\top}$ for all $g \in \prod_{a\in A} \mathsf{L}(a)$.

According to the definition (1) of the mapping ${}^{\perp}\colon \mathbf{P}(B) \to \prod_{a\in A} \mathsf{L}(a)$ we obtain for all $b \in B$ and all $a \in A$

$$\{b\}^{\perp}(a) = \bigwedge_{b'\in\{b\}} R(b', a).$$

Since the right side of this equality expresses the infimum over the one-element set $\{R(b, a)\}$ we obtain

$$\{b\}^{\perp}(a) = \bigwedge_{b'\in\{b\}} R(b', a) = R(b, a).$$

This yields that the value of the incidence relation $R(b, a)$ is fully determined by the a-th projection of $\{b\}^{\perp}$. Moreover, due to [6] the following assertion is valid:

Let B be a non-empty set and $\mathsf{L}(a)$ be a system of complete lattices. Then any two Galois connections (ϕ_1, ψ_1), (ϕ_2, ψ_2) between $\mathbf{P}(B)$ and $\prod_{a\in A} \mathsf{L}(a)$ are equal if and only if

$$\phi_1\big(\{b\}\big)(a) = \phi_2\big(\{b\}\big)(a)$$

for all $b \in B$ and for all $a \in A$.

Since we assume that particular projections of the elements of $\mathfrak{C}$ form dually isomorphic closure systems, we already know that there is a Galois connection (ϕ, ψ) between $\mathbf{P}(B)$ and $\prod_{a\in A} \mathsf{L}(a)$ such that the corresponding fixed points of (ϕ, ψ) form the lattice $\mathfrak{C}$. Hence we define the incidence relation $R\colon B \times A \to \bigcup_{a\in A} \mathsf{L}(a)$ as follows:

$$R(b, a) := \phi(\{b\})(a), \text{ for all } b \in B, a \in A.$$

From this definition it follows that

$$\phi(\{b\})(a) = R(b, a) = \{b\}^{\perp}(a)$$

for all $b \in B$, $a \in A$ and, due to the above-mentioned assertion, this yields ${}^{\perp} = \phi$ and ${}^{\top} = \psi$. In order to determine all the values $R(b, a)$ we must find the corresponding values of $\phi(\{b\})(a)$ in the generalized one-sided concept lattice $\mathfrak{C}$.

For this purpose we use the characterization of the one part of Galois connections $\phi\colon \mathbf{P}(B) \to \prod_{a\in A} \mathsf{L}(a)$ as the composition of the closure operator c on the set B and the dual isomorphism f between closure systems in $\mathbf{P}(B)$ and $\prod_{a\in A} \mathsf{L}(a)$ respectively. In this case $\phi = c \circ f$, i.e., $\phi(X) = f(c(X))$ for all $X \subseteq B$.

The isomorphism f is given directly by the ordered pairs in the generalized one-sided concept lattice $\mathfrak{C}$. If $(X, g) \in \mathfrak{C}$, then the dual isomorphism f is given by

$$f(X) = g, \text{ for all } X \subseteq B.$$

Hence the main goal is to determine the values $c(\{b\})$ for all $b \in B$.

From the definition of closure operator, it follows that for each $b \in B$ the value $c(\{b\})$ is the smallest subset appearing in $\mathfrak{C}$ which contains the given element b. For this reason it is convenient to deal with minimal elements of the concept lattice $\mathfrak{C}$. If (X, g) is a minimal element of $\mathfrak{C}$ then for all $b \in X$ it holds $c(\{b\}) = X$ and consequently

$$\phi(\{b\}) = f(c(\{b\})) = f(X) = g$$

for all $b \in X$. In the next step, we can remove this concept from the concept lattice $\mathfrak{C}$ and find another minimal element, say (X_1, g_1). Let us note that after removing any of the concepts (X, g) from $\mathfrak{C}$ the resulting structure in no longer a lattice in general. However it is still a partially ordered set, thus the notion of a minimal element can be used again. In this case $c(\{b\}) = X_1$ for all $b \in X_1 \smallsetminus X$. In this way we can proceed, until we exhaust all the elements in the object set B.

The whole procedure is described in more detail in Algorithm 2.

The correctness of this algorithm follows from the fact that for an object b and an attribute a the value $R(b, a)$ is uniquely determined by the value $\phi(\{b\})(a)$ where ϕ represents one part of the Galois connection (ϕ, ψ) corresponding to the input set $\mathfrak{C}$. Moreover $\phi(\{b\}) = \phi(X)$ where X is the closure of the element b in Galois connection (ϕ, ψ), i.e., $X = \psi(\phi(\{b\}))$. Consequently $\phi(\{b\}) = \phi(X) = g$ with $(X, g) \in \mathfrak{C}$. Closures of the one-element subsets are minimal with respect to the closure operator, hence in the **while** loop (line 2 - 13) the algorithm works with the minimal concepts in $\mathfrak{C}$. In the **for all** loop (line 5 - 12) the values $R(b, a)$ for $b \in X \smallsetminus S$, $a \in A$ are determined (**for all** loop (line 7 - 9)). Let $b \in B$ be an object, $(X, g) \in \mathfrak{C}$ be a minimal concept such that $b \in X$ and suppose that $b \notin S$, i.e., the values $R(b, a)$ for $a \in A$ are not determined yet. Then X is unique with this property and for all $a \in A$ the value $R(b, a)$ is determined correctly. By contrary assume that there is another subset X' with $(X', g') \in \mathfrak{C}$, $b \in X'$ and $X \nsubseteq X'$. Then $b \in X \cap X' \subsetneq X$ and $(X \cap X', g'') \in \mathfrak{C}$ for some $g'' \in \prod_{a\in A} \mathsf{L}(a)$ since the first components of $\mathfrak{C}$ form a closure system in $\mathbf{P}(B)$. This yields a contradiction to the fact that X is the minimal concept in $\mathfrak{C}$ for which $R(b, a)$ is not determined.

Finally, we describe the time complexity of Algorithm 2. Let $\mathfrak{C}$ be its input. Again, denote by n the number of

Algorithm 2 for finding the incidence relation R

```
Input: a generalized one-sided concept lattice $\mathfrak{C}$
Output: the incidence relation $R$
1:  $S \leftarrow \emptyset$    ▷ Set of objects $b$ for which $R(b,a)$ has already been determined
2:  while $S \neq B$ do    ▷ Repeat until all values are determined
3:      $m \leftarrow (X,g) : (X,g)$ a minimal element in $\mathfrak{C}$    ▷ Find a minimal concept
4:      $\mathfrak{C} \leftarrow \mathfrak{C} \smallsetminus \{m\}$    ▷ Remove the minimal concept $m$ from $\mathfrak{C}$
5:      for all $b \in X$ where $(X,g) = m$ do
6:          if $b \notin S$ then    ▷ $b$ has no value $R(b,a)$ yet
7:              for all $a \in A$ do
8:                  $R(b,a) \leftarrow g(a)$    ▷ Determination of the value $R(b,a)$
9:              end for
10:             $S \leftarrow S \cup \{b\}$    ▷ Add the object $b$ to the set $S$
11:         end if
12:     end for
13: end while
14: return $R$
```

concepts in $\mathfrak{C}$ and denote by k the number of all objects in the set B. In the worst case the **while** loop (line 2 - 13) has n repetitions (this happens when $\mathfrak{C}$ is a chain). A minimal concept of $\mathfrak{C}$ can be found in $O(|\mathfrak{C}|)$ time (see Algorithm 3).

Algorithm 3 for finding a minimal concept in $\mathfrak{C}$

```
Input: a generalized one-sided concept lattice $\mathfrak{C}$
Output: a minimal concept $m = (X,g)$
$m \leftarrow$ an arbitrary element in $\mathfrak{C}$
for all $m' \in \mathfrak{C}$ do
    if $m' < m$ then
        $m \leftarrow m'$
    end if
end for
return $m$
```

Since $X \subseteq B$ for all concepts $(X,g) \in \mathfrak{C}$, the **for all** loop (line 5 - 12) has at most k repetitions. Other loops can be executed in constant time, hence we obtain that the time complexity of Algorithm 2 is

$$\sum_{i=0}^{n-1} k \cdot c \cdot (n-i) = ck \cdot \frac{n \cdot (n+1)}{2} \in k \cdot O(n^2).$$

Since in many real situations $|B| \leq |\mathfrak{C}|$, we can conclude that the time complexity of Algorithm 2 is $O(n^3)$.

4 Conclusion

In this paper we have presented an algorithm for the inverse problem of generalized one-sided concept lattices, i.e., how to determine a generalized one-sided formal context from a given generalized one-sided concept lattice. This provides a possibility to express information about object-attribute models with different types of attributes in the form of hierarchical structures represented by generalized one-sided concept lattices.

Acknowledgement

The authors would like to thank the anonymous reviewer for the helpful and constructive comments which helped enhance the presentation of this paper.

This work was supported by the Slovak Research and Development Agency under contracts APVV-0035-10 and APVV-0482-11; by the Slovak VEGA Grants 2/0028/13, 1/0729/12 and 1/0497/11; by the ESF Fund CZ.1.07/2.3.00/30.0041.

References

[1] F. Babič, P. Bednár, F. Albert, J. Paralič, J. Bartók, L. Hluchý (2011) Meteorological phenomena forecast using data mining prediction methods, *Proceedings of 3rd International Conference on Computational Collective Intelligence, ICCCI 2011*, LNAI 6922, pp. 458–467.

[2] R. Bělohlávek (1999) Lattices generated by binary fuzzy relations., *Tatra Mountains Math. Publ.*, 16, pp. 11–19.

[3] R. Bělohlávek (2001) Lattices of Fixed Points of Fuzzy Galois Connections, *Math. Log. Quart.*, 47 (1), pp. 111–116.

[4] P. Butka, M. Sarnovský, P. Bednár (2008) One approach to combination of FCA-based local conceptual models for text analysis - Grid-based approach, *Proceedings of the 6th International Symposium on Applied Machine Intelligence and Informatics, SAMI 2008*, pp. 131–135.

[5] P. Butka, P. Bednár, F. Babič, K. Furdík, J. Paralič (2009) Distributed task-based execution engine for support of text-mining processes, *Proceedings of the 7th International Symposium on Applied Machine Intelligence and Informatics, SAMI 2009*, pp. 29–34.

[6] P. Butka, J. Pócs (2013) Generalization of one-sided concept lattices, *Computing and Informatics*, 32 (2), pp. 355–370.

[7] B. Ganter, R. Wille (1999) *Formal concept analysis, Mathematical foundations.*, Springer, Berlin.

[8] C. Havrilová, F. Babič (2013) Financial data analysis using suitable open-source Business Intelligence solutions, *Proceedings of the 11th International Symposium on Applied Machine Intelligence and Informatics, SAMI 2013*, pp. 257–262.

[9] A. Jaoua, S. Elloumi (2002) Galois connection, formal concepts and Galois lattice in real relations: application in a real classifier, *The Journal of Systems and Software*, 60, pp. 149–163.

[10] S. Krajči (2003) Cluster based efficient generation of fuzzy concepts, *Neural Network World*, 13 (5), pp. 521–530.

[11] J. Medina, M. Ojeda-Aciego, J. Ruiz-Calviño (2009) Formal concept analysis via multi-adjoint concept lattices, *Fuzzy Sets and Systems*, 160,pp. 130–144.

[12] O. Ore (1944) Galois Connexions, *Trans. Amer. Math. Soc.* 55, pp. 493–513.

[13] J. Pócs (2012) Note on generating fuzzy concept lattices via Galois connections, *Information Sciences*, 185 (1), pp. 128–136.

[14] M. Sarnovský, P. Butka, J. Paralič (2009) Grid-based support for different text mining tasks, *Acta Polytechnica Hungarica*, 6 (4), pp. 5–27.

[15] M. Sarnovský, P. Butka (2012), Cloud computing as a platform for distributed data analysis, *Proceedings of the 7th Workshop on Intelligent and Knowledge Oriented Technologies, Smolenice*, pp. 177–180.

Using Semantic Clustering for Detecting Bengali Multiword Expressions

Tanmoy Chakraborty
Department of Computer Science & Engineering
Indian Institute of Technology Kharagpur, India-721302
E-mail: its_tanmoy@cse.iitkgp.ernet.in; http://cse.iitkgp.ernet.in/ tanmoyc

Keywords: multiword expressions, compound noun, semantic clustering, statistical analysis

Received: July 12, 2013

Multiword Expressions (MWEs), a known nuisance for both linguistics and NLP, blur the lines between syntax and semantics. The semantic of a MWE cannot be expressed after combining the semantic of its constituents. In this study, we propose a novel approach called "semantic clustering" as an instrument for extracting the MWEs especially for resource constraint languages like Bengali. At the beginning, it tries to locate clusters of the synonymous noun tokens present in the document. These clusters in turn help measure the similarity between the constituent words of a potential candidate using a vector space model. Finally the judgment for the suitability of this phrase to be a MWE is carried out based on a predefined threshold. In this experiment, we apply the semantic clustering approach only for noun-noun bigram MWEs; however we believe that it can be extended to any types of MWEs. We compare our approach with the state-of-the-art statistical approach. The evaluation results show that the semantic clustering outperforms all other competing methods. As a byproduct of this experiment, we have started developing a standard lexicon in Bengali that serves as a productive Bengali linguistic thesaurus.

Povzetek: V prispevku je predstavljena metoda za semantično gručenje večbesednih izrazov.

1 Introduction

Over the past two decades or so, Multiword Expressions (MWEs) have been identified with an increasing amount of interest in the field of Computational linguistics and Natural Language Processing (NLP) [1]. The term "MWE" is used to refer to various types of linguistic units and expressions including idioms (kick the bucket, 'to die'), compound noun (village community), phrasal verbs (find out, 'search'), other habitual collocations like conjunctions (as well as), institutionalized phrases (many thanks) etc. However, while there is no universally agreed definition for MWE as yet, most researchers use the term to refer to those frequently occurring phrasal units which are subject to a certain level of semantic opaqueness, or non-compositionality. Sag et al. [30] defined them as "idiosyncratic interpretations that cross word boundaries (or spaces)."

MWEs are treated as a special case in semantics since individual components of an expression often fail to keep their meanings intact within the actual meaning of that expression. This opaqueness in meaning may be partial or total depending on the degree of compositionality of the whole expression [12]. MWEs have been studied for decades in Phraseology under the term "phraseological unit" [5]. But in the early 1990s, MWEs started receiving increasing attention in corpus-based computational linguistics and NLP. A number of research activities on MWEs have been carried out in various languages like English, German and many other European languages. Various statistical co-occurrence measurements like Mutual Information (MI) [15], Log-Likelihood [21], Salience [26] have been suggested for the identification of MWEs.

In the case of Indian languages, a considerable amount of research has been conducted in compound noun MWE extraction [28], complex predicate extraction [17], clustering based approach [12] and a classification based approach for identifying Noun-Verb collocations [33]. Bengali, one of the more important Indo-Iranian languages, is the sixth-most popular language in the world and spoken by a population that now exceeds 250 million[1]. Geographical Bengali-speaking population percentages are as follows: Bangladesh (over 95%), and the Indian States of Andaman & Nicobar Islands (26%), Assam (28%), Tripura (67%), and West Bengal (85%). The global total includes those which are spoken in the Diaspora in Canada, Malawi, Nepal, Pakistan, Saudi Arabia, Singapore, the United Arab Emirates, the United Kingdom, and the United States. In Bengali, works on automated extraction of MWEs are limited in number. One method of automatic extraction of Noun-Verb MWE in Bengali [2] has been carried out using morphological evidence and significance function. They have classified Bengali MWEs based on the morphosyntactic flexibilities and proposed a statistical approach for extracting the verbal compounds from a medium size corpus.

In this paper, we propose a framework for identifying MWEs from the perspective of semantic interpretation of

[1] http://en.wikipedia.org/wiki/Bengali_language

MWEs that the meanings of the components are totally or partially diminished in order to construct the actual semantics of the expression. A clustering technique is employed to group all nouns that are related to the meaning of the individual component of an expression. Two types of similarity techniques based on vector space model are adapted to make a binary classification (MWE or Non-MWE) of potentially candidate phrases. We hypothesize that the more similar the components of an expression, the less probable that their combination forms a MWE. We test our hypothesis on the noun-noun bigram phrases. We also illustrate the efficiency of our model after translating the individual components of a phrase in English and fed these components into the WordNet::Similarity module module – an open-source package developed at the University of Minnesota for calculating the lexical similarity between word (or sense) pairs based on variety of similarity measure. In this paper, we test our models with different cut-off values that define the threshold of (dis)similarity and the degree of compositionality of a candidate phrase. Experimental results corroborate our hypothesis that the dissimilarity of the meaning of constituent tokens enhances the chance of constructing a MWE. The use of English WordNet, quite strikingly, substantiates its enormous productivity in identifying MWEs from Bengali documents.

The remainder of this paper is organized as follows. Section 2 introduces a preliminary study about the Bengali MWEs and their morpho-syntactic based classification. Then the detailed description of candidate selection and the baseline system are described in section 3 and section 4 respectively. Section 5 illustrates traditional statistical methodologies for extracting MWEs from the document. Section 6 presents an elaborate description of semantic clustering approach. The introduction of English WordNetSimilarity in identifying Bengali MWEs is presented in section 7. The metrics used for evaluating the systems and experimental results are discussed in section 8. The discussion regarding the utilities and shortcomings of our model is illustrated in section 9 and the concluding part is drawn in section 10.

2 Multiword expressions (MWEs)

Though MWEs are understood quite easily and their acquisition presents no difficulty to native speakers (though it is usually not the case for second language learners), it is hard to identify what features distinguish MWEs from free word combinations. Concerning this issue, the following MWE properties are mentioned in the literature: reduced syntactic and semantic transparency; reduced or lack of compositionality; more or less frozen or fixed status; possible violation of some otherwise general syntactic patterns or rules; a high degree of lexicalization (depending on pragmatic factors); a high degree of conventionality [8].

No consensus exists so far on the definition of MWEs, but almost all formulations found in research papers emphasize the idiosyncratic nature of this linguistic phenomenon by indicating that MWEs are "idiosyncratic interpretations that cross word boundaries (or spaces)" [30]; "a sequence of words that acts as a single unit at some level of linguistic analysis, ... they are usually instances of well productive syntactic patterns which nevertheless exhibit a peculiar lexical behavior" [8]; "a MWE is composed of two or more words that together form a single unit of meaning, e.g., frying pan, take a stroll, and kick the bucket, ... Semantic idiosyncrasy, i.e., the overall meaning of a MWE diverges from the combined contribution of its constituent parts" [24].

2.1 Noun-Noun MWEs

In the past few years, noun compounds have been a constant source of concern to the researchers towards the goal of full text understanding [5, 7]. Compound nouns are nominal compounds where two or more nouns are combined to form a single phrase such as 'golf club' or 'computer science department' [5]. There is also a broader class of nominal MWEs where the modifiers are not restricted to be nominal, but can also be verbs (e.g., hired help) or adjectives (e.g., open secret). To avoid confusion in this article, we will use the term compound nouns when referring to this broader class, throughout the paper, we term this broader class.

Compound noun MWEs can be defined as a lexical unit made up of two or more elements, each of which can function as a lexeme independent of the other(s) when they occur separately in different contexts of the document. The combination of these constituents shows some phonological and/or grammatical isolation from their normal syntactic usages. One property of compound noun MWEs is their underspecified semantics. For example, while sharing the same "head noun" (i.e., rightmost noun in the noun compound), there is less semantic commonality between the components such as 'nut tree', 'cloths tree' and 'family tree' [5]. In each case, the meaning of the compound nouns relates to a sense of both the head and the modifier, but the precise relationship is highly varied and not represented explicitly in any way. Noun-Noun (NN) compounds are the subset of the compound nouns consisting of two consecutive nouns side by side. In English, NN compounds occur in general with high frequency and high lexical and semantic variabilities. A summary examination of the 90 million word written component of the British National Corpus unearthed over 400,000 NN compound types, with a combined token frequency of 1.3 million; that is, over 1% of words in the BNC are NN compounds [32].

In Bengali, similar observations are noticed when dealing with the various types of multiword expressions like compound nouns (taser ghar, 'house of cards', 'fragile'), complex predicates such as conjunct verbs (*anuvab kara*, 'to feel') and compound verbs (*uthe para*, 'to arise'), idioms (*matir manus*, 'down to the earth'), named-entities (*Rabindranath Thakur*, 'Rabindranath Tagore') etc. Bengali is a language consisting of high morpho-syntactic variation at the surface level. The use of NN multiword expressions in Bengali is quite common. For example, NN compounds especially, idioms (*taser ghar*, 'fragile'), institutionalized phrases (*ranna ghar*, 'kitchen'), named-entities (*Rabindranath Thakur*, 'Rabindranath Tagore'), numbers (*panchso noi*, 'five hundred and nine'), kin terms (*pistuto bhai*, 'maternal cousin') etc. are very frequently used in Bengali literature. In the next subsection, we classify the compound nouns occurred in Bengali based on their morpho-syntactic properties.

2.2 Classifications of Bengali compound noun MWEs

Compound noun MWEs can occur in open (components are separated by space(s)), closed (components are melded together) or hyphenated forms (components are separated by hyphen(s)), and satisfy semantic non-compositionality, statistical co-occurrence or literal phenomena [28] etc. Agarwal et al. (2004) classified the Bengali MWEs in three main classes using subclasses. Instead, we propose seven broad classes of Bengali compound noun MWEs considering their morpho-syntactic flexibilities, as follows:

- **Named-Entities (NE):** Names of people (*Rabindranath Thakur*, 'Rabindranath Tagore'), names of locations (*Bharat-barsa*, 'India'), names of organizations (it Pashchim Banga Siksha Samsad, 'West Bengal Board of Education') etc. where inflection is only allowed to be added to the last word.

- **Idiomatic Compound Nouns:** These are non-productive[2] and idiomatic in nature, and inflection can be added only to the last word. The formation of this type is due to the hidden conjunction between the components or absence of inflection from the first component (*maa-baba*, 'mother and father').

- **Idioms:** They are also compound nouns with idiosyncratic meaning, but the first noun is generally in the possessive form (*taser ghar*, 'fragile'). Sometimes, individual components may not carry any significant meaning and may not represent a valid word (*gadai laskari chal*, 'indolent habit'). For them, no inflection is allowed even to the last word.

- **Numbers:** They are highly productive, impenetrable and allow slight syntactic variations like inflections. Inflection can be added only to the last component (*soya sat ghanta*, 'seven hours and fifteen minutes').

- **Relational Noun Compounds:** They are mainly kin terms and consist mostly of two tokens. Inflection can be added to the last word *pistuto bhai*, 'maternal cousin').

- **Conventionalized Phrases:** Sometimes, they are called as 'Institutionalized phrases'. Although not necessarily idiomatic, a particular word combination coming to be used to refer to a given object. They are productive and have unexpectedly low frequency and in doing so, contrastively highlight the statistical idiomaticity of the target expression (*bibhha barshiki*, 'marriage anniversary'). Simile Terms: They are analogy term in Bengali and sometime similar to the idioms except that they are semi-productive (*hater panch*, 'remaining resource').

- **Reduplicated Terms:** Reduplications are non-productive and tagged as noun phrases. They are further classified as onomatopoeic expressions (*khat khat*, 'knocking'), complete reduplication (*bara-bara*, 'big big'), partial reduplication (*thakur-thukur*, 'God'), semantic reduplication (*matha-mundu*, 'head'), correlative reduplication (*maramari*, 'fighting') [11].

Identification of reduplication has already been carried out using the clues of Bengali morphological patterns [11]. A number of research activities in Bengali Named Entity (NE) detection have been conducted [23], but the lack of publicly available standard tools to detect NEs inhibits the incorporation of them within the existing system. Therefore, we discard the identification of NEs from this experiment. Kin terms and numbers can be easily captured by some well-developed lexicons because they are small in number and form a closed set in Bengali [2]. The present work mainly focuses on the extraction of productive and semi-productive bigrams, compound noun MWEs like idioms, idiomatic compound nouns, and simile terms (which are in open or hyphenated form) from a document using a semantic clustering technique.

3 Semi-automated approach for candidate extraction

3.1 Corpus acquisition and bigram extraction

Resource acquisition is one of the challenging obstacles to work with electronically resource constrained languages like Bengali. However, we crawled a large number of Bengali articles written by the noted Indian Nobel laureate Rabindranath Tagore[3]. While we are primarily interested in token level or phrase level characteristics, document information (e.g., the order of the documents, variation of the size of the documents, length normalization etc.) has not been maintained and manipulated in the experiment. Therefore, we merged all the articles and prepared a raw corpus consisting of 393,985 tokens and 283,533 types. The actual motivation for choosing the literature domain

[2] A phrase is said to be "productive" if new phrases can be formed from the combinations of syntactically and semantically similar component words of the original phrase.

[3] `http://www.rabindra-rachanabali.nltr.org`

in the present task was to obtain useful statistics to further help Stylometry analysis [9]. However in literature, the use of MWEs is greater than in the other domains like tourism, newspapers, scientific documents etc. because the semantic variability of MWEs offers writers more expressive terms. In Bengali literature, idiomatic expressions and relations terms are quite frequently used.

Since the preliminary crawled corpus was noisy and unformatted, we used a basic semi-automatic pre-processing technique to make the corpus suitable for parsing. We used a Bengali shallow parser[4] to identify the POS, chunk, root, inflection and other morphological information of each token. We observed that some of the tokens were misspelled due to typographic and phonetic errors. Thus, the Shallow parser could not be able to detect the actual root and inflection of these two variations. To make the system fully automated, we allowed retaining the types of variations into the cleaned text.

After pre-processing, bigram noun sequences whose constituents were in the same chunk were extracted using their POS and chunk categories. We observed that during the parsing phase, the Shallow parser could not disambiguate common nouns ('NN') and proper nouns ('NNP') appropriately. The reason could be the continuous need to coin new terms for new concepts. We took both of them and manually filtered the named-entities from the collected list so that we could accumulate most of the proper nouns for our main experimental module. Although the chunk information helps to identify the boundary of a phrase, some of the phrases belong to chunks having more than two nouns. The frequency of these phrases is also identified during the evaluation phase. Now, a bigram nominal candidate phrase can be thought of as $< M1\ M2 >$. The morphological heuristics used to separate the candidates are described in Table 1. After the first phase, a list of possible candidates was collected which was fed into the annotation phase.

Heuristics	
POS	POS of each bigram must be either 'NN' or 'NNP'
Chunk	M1 and M2 must be in the same 'NP' chunk
Inflection	Inflection[5] of $M1$ must be 'null', (-r), (-er), (-e), (-y) or (-yr)

Table 1: Heuristics for the candidate selection

3.2 Annotation study

Three anonymous annotators – linguistic experts working on our project – were hired to carry out the annotation. They were asked to divide all extracted phrases into four classes and definitions of the classes using the following definitions:

Class 1: Valid NN MWEs (M): phrases which show total non-compositionality and their meanings are hard to predict from their constituents; e.g., *hater panch* ('remaining resource').

Class 2: Valid NN semantic collocations but not MWEs (S): phrases which exhibit partial or total compositionality (e.g., act as institutionalized phrases) and show statistical idiomaticity; e.g., *bibaha barsiki* ('marriage anniversary').

Class 3: Invalid collocations (B): phrases enlisted due to bigrams in an n-gram chunk having more than two components; e.g., *porbot sohorer*, ('of mountain town').

Class 4: Invalid candidates (E): phrases enlisted due to the error in parsing like POS, chunk, inflection including named-entities; e.g., *granthagar tairi* ('build library').

Class 3 and class 4 types were filtered initially and their individual frequencies are noted as 24.37% and 29.53% respectively. Then the remaining 46.10% (628 phrases) of the total candidates were annotated and labeled as MWE (M) or S (Semantically collocated phrases), and they were fed into the evaluation phase. We plan to make the dataset publicly available soon.

The annotation agreement was measured using standard Cohen's kappa coefficient (κ) [16]. It is a statistical measure of inter-annotation agreement for qualitative (categorical) items. It measures the agreement between two raters who separately classify items into some mutually exclusive categories. We employ another strategy in addition with kappa (κ) to calculate the agreement between annotators. We choose the measure of agreement on set-valued items ($MASI$) [29] that is used for measuring agreement in the semantic and pragmatic annotations. $MASI$ is a distance between sets whose value is 1 for identical sets, and 0 for disjoint sets. For sets A and B, it is defined as: $MASI = J * M$, where the Jaccard metric (J) is:

$$J = \frac{A \cap B}{A \cup B} \tag{1}$$

Monotonicity (M) is defined as follows:

$$M = \begin{cases} 1, & if\ A = B \\ 2/3, & if\ A \subset B\ or\ B \subset A \\ 1/3, & if\ A \cap B \neq \phi, A - B \neq \phi\ \&\ B - A \neq \phi \\ 0, & if\ A \cap B = \phi \end{cases} \tag{2}$$

The inter-annotation agreement scores of three annotators are presented in Table 2. Among the 628 types of noun-noun candidates, half of them selected randomly were used in the development phase and the remaining were used in the testing phase.

4 Baseline system

As mentioned earlier, the task of identifying Bengali compound nouns from a document has had little attention in the literature, and thus there is no prior developed methodology that can be used for the baseline. Therefore, in this experiment, we simply adapt a heuristic to develop our baseline system. The phrases which do not affix any nominal chunk and determinant at the prefix and suffix positions are selected as MWEs in the baseline system. The baseline system naturally reaches high accuracy in terms of recall since most of the identified MWEs satisfy the heuristics mentioned above. But in terms of precision, it shows very

[4] http://ltrc.iiit.ac.in/analyzer/bengali

MWEs [# 628]	Agreement between pair of annotators			
	A1-A2	A2-A3	A1-A3	Average
KAPPA	87.23	86.14	88.78	87.38
MASI	87.17	87.02	89.02	87.73

Table 2: Inter-annotation agreement

low accuracy (38.68%) since many collocated and fully-compositional elements were wrongly identified as MWEs. The main challenge of our model was to filter these irrelevant collocations from the selected candidate set.

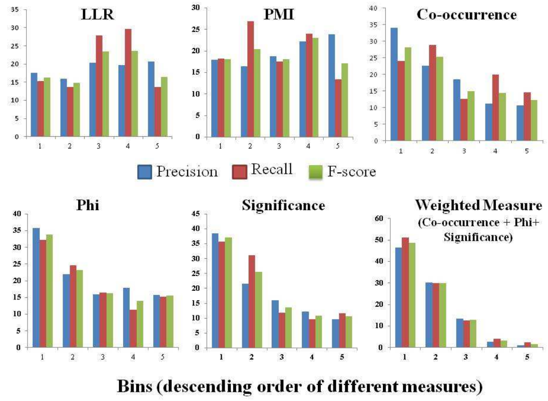


Figure 1: Performance of all the statistical measures and the weighted measure.

5 Statistical methodologies

We started our experiment with the traditional methodology of collocation detection. Previous literature [15] [21] [26] shows that various statistical methodologies could be incorporated in identifying MWEs from a large corpus. In this experiment, we developed a statistical system using these previous techniques and modified them according to our requirements[6]. It is worth noting that frequency information of the candidate phrases in a corpus is a strong clue for labeling them as MWEs since it provides the evidence of more certainty of occurrence than randomness. However, for a resource-constrained language like Bengali, infrequent occurrence of candidates may not give any reasonable conclusion to judge them as MWEs (or Non-MWEs) because the size of the corpus itself is generally not adequate for statistical analysis. Therefore, instead of taking the frequency information directly, we took five standard association measures namely Point-wise Mutual Information (PMI) [15], Log-Likelihood ratio (LLR) [21], Co-occurrence measure [2], Phi-coefficient and Significance function [2] for extracting NN Multiword Expressions. A combined weighted measurement is proposed for the identification task, which is helpful to compute bigram collocation statistics. We ranked the list individually based on each of the statistical measures. We noticed in the comparative study that the results obtained by the frequency-based statistics like PMI and LLR could not identify MWEs at the top position of the ranked list. Therefore, we posited that the lexico-semantic affinity among the constituents could unleash the dependency of frequency information in the measurement. Final evaluation combined all the statistical features mentioned above. Experimental results on the development dataset show that Phi-coefficient, Co-occurrence and Significance functions which are actually based on the principle of collocation produce more accurate results compared to direct frequency-based measurements like LLR, PMI in the higher ranks. So, these three measures are considered in the weighted scheme to assign certain weights to the candidate phrases. After a continuous weight tuning over the development data, the best weights for Co-occurrence, Phi and Significance functions are reported as 0.45, 0.35 and 0.20 respectively for the combined measurement. The individual score of each measure is normalized before assigning weights so that they fall in the range of 0 to 1. For each measurement, the scores have been sorted in descending order and the total range is divided into five bins (bin 1 signifies higher ranked bin). The intuition is that the more the value of the statistical measure for a candidate phrase, the more it behaves like a MWE. The metrics used to evaluate the statistical systems are described below:

Precision in bin i (P_i) = (Number of MWEs present in the ith bins) / (total number of candidates in ith bins)

Recall in bin i (R_i) = (Number of MWEs present in the ith bins) (total number of MWEs in the documents)

F-score in bin i (F_i) = $(2 * P_i * R_i)(P_i + R_i)$

Figure 1 shows the results obtained from five association measures and the combined weighted measures over the test dataset.

6 Semantic clustering approach

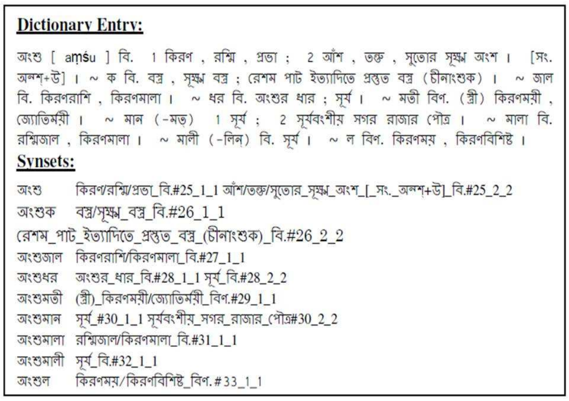
Multiword Expressions represent a core semantic issue that can be partially resolved by morphological or statistical clues. However, it often fails at capturing the underlying semantic notion of forming a new multiword expression, i.e., the meaning of the entire expression cannot be predicted by aggregating the meaning of its components. Our proposed approach aims to handle these drawbacks by considering individual senses induced by the components of an expression. This approach tries to cluster semantically related words present in the document. However,

[6] Interested readers are encouraged to go through the research dissertation by [10]

for a particular token present in the document, finding semantically similar words appeared in the corpus can be carried out by looking at the surroundings tokens and finding the synonymous entries of the surrounding words within a fixed context window. However in that case, a high number of occurrences of a particular token should be needed in a corpus in order to obtain statistically significant evidences. Therefore, in a medium-size corpus, it is hard to extract the cluster of synonyms. Since the electronic resources such as newspapers, weblogs may not be present for all the languages and the presence of frequent MWEs in such contents are rare, we focus on extracting the MWEs only from the medium size crawled corpus. However, semantics of a word may be obtained by analyzing its similarity set called the synset that indeed expresses its meaning in different contexts. Therefore, semantic distance of two tokens in a phrase can be measured by comparing their synsets properly. Higher value of the similarity between two sets indicates semantic closeness of two tokens to each other. For instance, let $M1$ and $M2$ be two components of a bigram $< M1\ M2 >$. For each component of the expression, semantically related words present in the documents are extracted by using the formatted Bengali monolingual dictionary (discussed in Section 6.1) and two separate clusters are formed for two tokens. Now, intersection of two clusters can be a suitable measure to judge the commonality of two components appeared in a bigram. Using these common elements, three different similarity measurements are proposed in our algorithm and the results are reported separately in Table 5 later. Finally, based on a predefined threshold, the candidate phrases were labeled as MWE or Non-MWE.

6.1 Restructuring the Bengali monolingual dictionary

To the best of our knowledge, no full-fledged WordNet or thesaurus is available in Bengali. In this section, we describe the construction of a Bengali thesaurus that aims not only to develop Bengali WordNet but also to identify the meaning of multiword expressions. Focusing mainly on MWEs, the present natural language resource is being developed from the available Bengali-to-Bengali monolingual dictionary (Samsada Bengali Abhidhana[7]). The monolingual dictionary contains each word with its parts-of-speech (Noun, Adjective, Pronoun, Indeclinable, Verb), phonetics and synonym sets. Synonym sets are separated using distinguishable notations based on similar or differential meaning. Synonyms of different sense with respect to a word entry are distinguished by a semicolon (;), and synonyms having same sense are separated by a comma (,). An automatic technique is devised to identify the synsets for a particular word entry based on the clues (, and ;) of similar and differential senses. The symbol tilde (~) indicates that the suffix string followed by the tilde (~) notation makes another new word concatenating with the original entry word. A snapshot of the modified synset entries of the Bengali word *Angshu* is shown in Figure 2. Table 3 shows the frequencies of different synsets according to their part-of-speech.



Dictionary Entry:

অংশু [aṃśu] বি. 1 কিরণ , রশ্মি , প্রভা ; 2 আঁশ , তন্তু , সুতোর সূক্ষ্ম অংশ । [সং. অন্শ্+উ] । ~ ক বি. বস্ত্র , সূক্ষ্ম বস্ত্র ; রেশম পাট ইত্যাদিতে প্রস্তুত বস্ত্র (চীনাংশুক) । ~ জাল বি. কিরণরাশি , কিরণমালা । ~ ধর বি. অংশুর ধার ; সূর্য । ~ মতী বিণ. (স্ত্রী) কিরণময়ী , জ্যোতির্ময়ী । ~ মান (-মত্) 1 সূর্য ; 2 সূর্যবংশীয় সগর রাজার পৌত্র । ~ মালা বি. রশ্মিজাল , কিরণমালা । ~ মালী (-লিন্) বি. সূর্য । ~ ল বিণ. কিরণময় , কিরণবিশিষ্ট ।

Synsets:

অংশু কিরণ/রশ্মি/প্রভা_বি.#25_1_1 আঁশ/তন্তু/সুতোর_সূক্ষ্ম_অংশ_[_সং._অন্শ্+উ]_বি.#25_2_2
অংশুক বস্ত্র/সূক্ষ্ম_বস্ত্র_বি.#26_1_1
রেশম_পাট_ইত্যাদিতে_প্রস্তুত_বস্ত্র_(চীনাংশুক)_বি.#26_2_2
অংশুজাল কিরণরাশি/কিরণমালা_বি.#27_1_1
অংশুধর অংশুর_ধার_বি.#28_1_1 সূর্য_বি.#28_2_2
অংশুমতী (স্ত্রী)_কিরণময়ী/জ্যোতির্ময়ী_বিণ.#29_1_1
অংশুমান সূর্য_#30_1_1 সূর্যবংশীয়_সগর_রাজার_পৌত্র#30_2_2
অংশুমালা রশ্মিজাল/কিরণমালা_বি.#31_1_1
অংশুমালী সূর্য_বি.#32_1_1
অংশুল কিরণময়/কিরণবিশিষ্ট_বিণ.#33_1_1

Figure 2: Monolingual dictionary entry and built synsets for the word *Angshu*.

6.2 Generating synsets of nouns

At the beginning of the clustering method, we generate a synonym set for each noun present in the corpus using the modified dictionary. However, the formatted dictionary can be assumed to be a close set of word entries $W^1, W^2, W^3, ..., W^m$ where the synsets of the entries look like:

$$W^1 = n_1^1, n_2^1, n_3^1, ... = n^1$$

$$W^2 = n_1^2, n_2^2, n_3^2, \dots = n^2$$

$$\dots$$

$$W^m = n_1^m, n_2^m, n_3^m, \dots = n^m$$

where $W^1, W^2, ..., W^m$ are the dictionary entries and n^i denotes the set of synsets of the entry W^i. Now each noun entry identified by the shallow parser in the document is searched in the synset entries of the dictionary for its individual existence with or without inflection. For instance, N is a noun in the corpus and it is present in the synsets of W^1, W^3 and W^5. Therefore, they become entries of the synset of N. Formally, this can be represented as follows.

$$Synset(N) = \{W^l, W^3, W^5\} \quad (3)$$

Equation 2 states that since the given noun N is present in the synsets of W^1, W^3 and W^5, the sense of these three dictionary entries are somehow related to the sense of N. Following this, the synonym noun tokens for each of the nouns present in the corpus are extracted from the dictionary. In short, the formatted dictionary indeed helps us cluster synonymous tokens corresponding to a particular noun present in a document.

[7] http://dsal.uchicago.edu/dictionaries/biswas-bangala/

Word Entries	Synset	Noun	Adjective	Pronoun	Indeclinable	Verb
47949	63403	28485	11023	235	497	1709

Table 3: Frequency information of the synsets with different part-of-speeches.

6.3 Semantic relatedness between noun synsets

The next task is to identify the similarity between the synsets of two nouns that can help measure the semantic relatedness between them. This is done by taking the intersection of the synsets and assigning a score to each such noun-pair to indicate the semantic affinity between two nouns. For instance, if N_i and N_j are two nouns in the document, and S_i and S_j are their corresponding synsets extracted using the technique stated in Section 6.2, then the commonality of the two nouns can be defined as:

$$Comm(N_i, N_j) = |S_i \cap S_j| \tag{4}$$

The above equation shows that the commonality is maximum when the similarity is measured with itself (i.e., $Comm(N_i, N_j)$ is maximum when $i = j$).

6.4 Semantic clustering of nouns

Using the scores obtained by the semantic commonality measure discussed in the previous subsection, we can build a cluster centered on a given noun present in the document such that the cluster constitutes all the nouns semantically related to the given noun (discussed in subsections 6.2 and 6.3). A score is assigned to each such noun present in the cluster representing the semantic similarity (discussed in subsection 6.3) between this noun and the noun present at the center of the cluster. An illustrative example is shown in Figure 3. For example, suppose the nouns identified by the Shallow parser in the document are

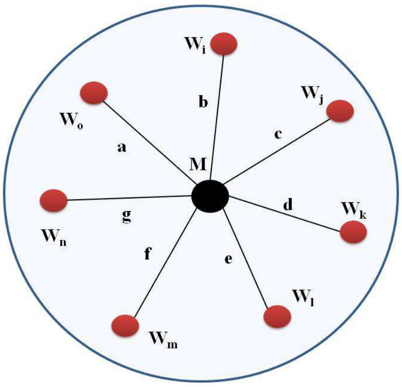
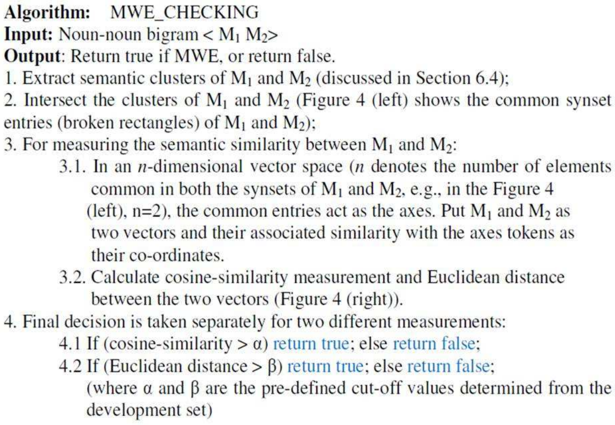


Figure 3: Semantic clustering for a given noun M and the associated commonality scores with similar nouns (W_i, W_j,..., W_o etc). In this figure, the semantic similarities of M with the other nouns are denoted by the weights (i.e., a, b, c etc.) of the edges.

6.5 Decision algorithm for identifying MWEs

We extract the candidates eligible for judging MWE in section 3. The elaborated algorithm to identify a noun-noun bigram (say, $< M1\ M2 >$) as MWEs is discussed below with an illustrative example shown in Figure 3.



```
Algorithm: MWE_CHECKING
Input: Noun-noun bigram < $M_1$ $M_2$>
Output: Return true if MWE, or return false.
1. Extract semantic clusters of $M_1$ and $M_2$ (discussed in Section 6.4);
2. Intersect the clusters of $M_1$ and $M_2$ (Figure 4 (left) shows the common synset
   entries (broken rectangles) of $M_1$ and $M_2$);
3. For measuring the semantic similarity between $M_1$ and $M_2$:
        3.1. In an $n$-dimensional vector space ($n$ denotes the number of elements
             common in both the synsets of $M_1$ and $M_2$, e.g., in the Figure 4
             (left), n=2), the common entries act as the axes. Put $M_1$ and $M_2$ as
             two vectors and their associated similarity with the axes tokens as
             their co-ordinates.
        3.2. Calculate cosine-similarity measurement and Euclidean distance
             between the two vectors (Figure 4 (right)).
4. Final decision is taken separately for two different measurements:
        4.1 If (cosine-similarity > $\alpha$) return true; else return false;
        4.2 If (Euclidean distance > $\beta$) return true; else return false;
            (where $\alpha$ and $\beta$ are the pre-defined cut-off values determined from the
            development set)
```

Here, we elaborate step 3 and step 4 since the central theme of the algorithm lies in these two steps. After identifying the common terms from the synsets of the components of a candidate, a vector space model is used to identify the similarity between the two components. In n-dimensional vector space, these common elements denote the axes and each candidate acts as a point in the n-dimensional space. The coordinate position of the point (each component of the candidate bigram) in each direction is represented by the similarity measure between the synsets of each component and the noun representing the axis in that direction. The cut-off value for the classification of a given candidate as MWE (or Non-MWE) is determined from the development dataset after several tries to get the best performance (described in step 4). We have seen significant results for the cut-off values (0.4-0.6) on the development set based on F-sore measure. Therefore, we report the results on the test dataset for each of these cut-off values separately in Table 4. In the experiment, we observe that the bigrams that are actual MWEs, mainly non-compositional phrases, show a low similarity score between the synsets of their components.

If we take an example of the Bengali idiom – *hater panch* ('remaining resource'), we can see that English WordNet defines two components of the idiom in the following way: *hat* ('hand') as 'a part of a limb that is farthest from the torso' and *panch* ('five') as 'a number which is one more than four'. So from these two glosses it is quite evident

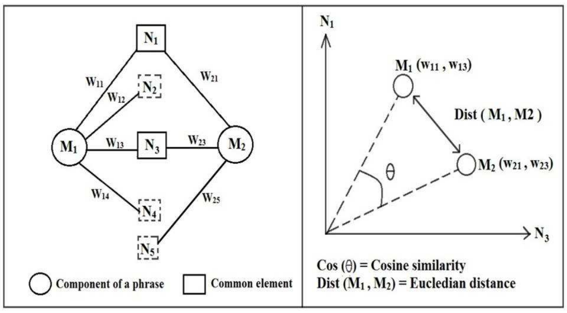


Figure 4: Intersection between the clusters of the components of a candidate bigram (left) and the similarity between two components (right)

that they are not at all semantically related. The synonym sets for these two components extracted from the formatted monolingual dictionary are as follows.

Synset (হাত) = { হস্ত, কর, পাণি, বাহু, ভুজ, কৌশল, হস্তক্ষেপ, ধারণ, রেখা, লিখিত, হস্তাক্ষর, হস্তান্তর, হাজা }
Synset (পাঁচ) = {পঞ্চ, সংখ্যা, কর্ম, গঙ্গা, গব্য, কন্যা, গুণ, গৌড়, তন্ত্র, তীর্থ, পঞ্চত্ব, পনেরো, পূর্ণিমা, পঞ্চাশ }

We can observe that the two synonym sets have no element in common and therefore their similarity score would be zero. In this case, a vector space model cannot be drawn in zero dimensional space. For them, a final concession weight is assigned to treat them as fully non-compositional phrases. To identify their non-compositionality, we need to show that their occurrences are not by mistake (i.e., because of a typo or due to unawareness of the author); rather they can occur side by side in several instances. But the concrete statistical proof can only be obtained using a large corpus. Here, for the candidate phrases which have zero similarity, we observe their existence more than one time in the corpus and then treat them as MWEs.

7 WordNet similarity measurement

We also incorporate English WordNet 2.1[8] in this experiment to measure the semantic distance between two Bengali words after translating them into English. Though the idea is trivial considering the manual intervention of the translation process, our main focus was to get an idea of how the semantic similarity of two components can help identify the combination as an MWE, and how a well-defined lexical tool is essential in the presently adapted linguistic environment. As already mentioned, WordNet::Similarity is an open-source package developed at the University of Minnesota for calculating the lexical similarity between word (or sense). Basically, it provides six measures of similarity and three measures of relatedness based on the WordNet lexical database [25]. The measures are based on the analysis of the WordNet hierarchy.

The measures of similarity are divided into two groups: path-based and information content-based. We chose two similarity measures in WordNet::Similarity for our experiments: WUP and LCH; WUP finds the path length to the root node from the least common subsumer (LCS) of the two word senses that is the most specific word sense they share as an ancestor [34]. In this experiment, we first translate the root of two Bengali components in a candidate phrase into their English forms using the Bengali-to-English Bilingual Dictionary[9]. Then these two words are run through the WordNet based Similarity module for measuring their semantic distance. A predefined cut-off value (μ) is determined from the development set to distinguish between an MWE and a simple compositional term. If the measured distance is less than the threshold, the similarity between them is less. The results are noted for different cut-off values as shown in Table 5. The bold font in each column shows the highest accuracy among different cut-off values.

8 Experimental results

We used standard IR metrics, i.e., Precision, Recall and F-score to evaluate the final results obtained from three similarity measuring modules (i.e., cosine-similarity, Euclidean distance and WordNet similairty) as discussed in the previous section. The evaluation of the systems was carried out on the previously mentioned hand-annotated dataset and the final results are shown in Table 5. The predefined threshold acquired from the development set was tuned to obtain the best results for all the similarity measures. Increasing recall with the increase of cut-off values indicates that most of the MWEs are identified across the wider threshold range. But the precision is not increasing gradually with the threshold. This result signifies that besides capturing most of the significant MWEs, it also considers more false positives at higher cut-off values. Our goal is to pick up an optimal point where both precision and recall stabilize with the reasonable results and minimize the erroneous predictions. The Cosine-similarity [**?**] achieves maximum precision at 0.5, whereas Euclidean distance and WordNet::Similarity achieve maximum precision at 0.4 and 0.5 respectively. The effect of English WordNet in identifying Bengali MWEs is noticeable in Table 5. Wordnet::Similarity identifies the maximum number of MWEs correctly at the cut-off of 0.5. Baldwin et al. (2003) suggested that WordNet::Similarity measure can be used to identify Multiword Expression decomposability. This is once again effective for Bengali MWE identification. There are also candidates with very low value of similarity between their constituents (e.g., *ganer jagat* (earth of song, 'affectionate of song')), yet they are discarded from

[8] http://www.d.umn.edu/~tpederse/similarity.html

[9] http://dsal.uchicago.edu/dictionaries/biswas-bengali/

Cut-off	Cosine-Similarity			Euclidean Distance			WordNet::Similarity		
	P	R	F	P	R	F	P	R	F
0.6	70.75	64.87	67.68	70.57	62.23	66.14	74.60	61.78	67.58
0.5	78.56	59.45	67.74	72.97	58.79	65.12	80.90	58.75	68.06
0.4	73.23	56.97	64.08	79.78	53.03	63.71	75.09	52.27	61.63

Table 4: Precision, Recall and F-score of three measures (in %) in clustering approach and WordNet::similarity measure.

this experiment because of their low frequency of occurrence in the corpus that could not reveal enough evidence of considering them as MWEs. Whether such an unexpectedly low frequent high-decomposable elements warrant an entry in the lexicon depends on the type of lexicon being built [4].

9 Discussion

At the beginning of the article, we claimed that the increasing degree of semantic similarity between two constituents of a candidate bigram indicates the less probability of the candidate to be a multiword expression. The statistical methodologies achieve low accuracy because the medium size corpus fails to unfold significant clue of their occurrences to label the non-compositional phrase as MWEs. We have adopted an approach taking into account the semantic interpretation of MWE that seems to be unconventional in the task of identifying MWEs in any language. In the experimental results, the semantic clustering approach outperforms the other systems. However, the clustering algorithm is able to identify those MWEs whose semantics are fully opaque from the semantics of their constituents (strictly non-compositional). But MWEs show a continuum spectrum from fully-compositional (e.g., idioms) to institutionalized phrases (e.g., traffic signal) where high statistical occurrence is the only clue to identify them as MWEs. These partial or transparent expressions are not captured by our system because of the lack of a large size standard corpus. The presence of the monolingual dictionary is another important criterion to carry out the proposed approach. It acts as a proxy for an individual noun to cumulate the related noun tokens. This algorithm assumes that every language should possess its own dictionary since it is the first and fundamental resource used not only for experimental purposes but also for language generation and understanding.

10 Conclusion

We hypothesized that sense induction using synonym set can assist in identifying multiword expressions in Bengali. We introduced a semi-automated approach to establish the hypothesis. We compared our results with the baseline system and the traditional statistical systems. We have shown that clustering measure can be an effective measure to enhance the extraction task of MWEs. The contributions of the paper are fourfold: firstly, we provide an efficient way of clustering noun tokens having similar sense; secondly, we propose a semantic similarity based approach for identifying MWEs; thirdly, it a preliminary attempt to reconstruct a Bengali monolingual dictionary as a standard lexical thesaurus and finally, the present task is a pioneering step towards the development of Bengali WordNet. At last, we would like to stress that this entire methodology can be used to identify MWEs in any other language domain. In the future, we plan to extend the algorithm to support all ranges of compositionality of Bengali MWEs. Moreover, we modify the semantic interpretation of MWEs to enlist partial and compositional phrases as much as possible. Furthermore, incorporating the Named-Entity recognizer can help develop a full-fledged MWE identification system. Finally, we will make the formatted monolingual dictionary publicly available soon and incorporate the strictly non-compositional MWEs which rarely occur in the medium-size corpus into the dictionary so that they are directly captured from the thesaurus.

References

[1] Rayson, P., Piao, S., Sharoff, S., Evert, S., Moriron, B. V. 2010. Multiword expressions: hard going or plain sailing? Language Resources and Evaluation, vol. 44, pp. 1–5.

[2] Agarwal, A., Ray, B., Choudhury, M., Sarkar, S., Basu, A. (2004). Automatic Extraction of Multiword Expressions in Bengali: An Approach for Miserly Resource Scenario. *In Proc. of International Conference on Natural Language Processing (ICON)*, pp. 165-174.

[3] Agirree, E., Aldezabal, I., Pociello, E. (2006). Lexicalization and multiword expressions in the Basque WordNet. *In Proc. of Third International WordNet Conference*. Jeju Island (Korea).

[4] Baldwin, T., Bannard, C., Tanaka, T., Widdows, D. (2003). An Empirical Model of Multiword Expression Decomposability. *In Proc. of the Association for Computational Linguistics, Workshop on Multiword Expressions: Analysis, Acquisition and Treatment*, Sapporo, Japan, pp. 89–96.

[5] Baldwin, T., Kim, S. N. (2010). Multiword Expressions, in Nitin Indurkhya and Fred J. Damerau

(eds.) *Handbook of Natural Language Processing*, Second Edition, CRC Press, Boca Raton, USA, pp. 267—292.

[6] Benson, M. (1990). Collocations and general-purpose dictionaries. *International Journal of Lexicography*, 3(1). pp. 23-35.

[7] Bhtnariu, C., Kim, S.N., Nakov, P., Oseaghdha, D., Szpakowicz., S., Veale, T. (2009). SemEval-2010 Task 9: The Interpretation of Noun Compounds Using Paraphrasing Verbs and Prepositions. *In Proc. of the NAACL Workshop on Semantic Evaluations: Recent Achievements and Future Directions (SEW-2009) at NAACL*, pp. 100-105.

[8] Calzolari, N., Fillmore, C., Grishman, R., Ide, N., Lenci, A., Macleod, C., Zampolli, A. (2002). Towards Best Practice for Multiword Expressions in Computational Lexicons. *In proc. of the Third International Conference on Language Resources and Evaluation (LREC)*, pp. 1934–1940.

[9] Chakraborty, T. (2012). Authorship Identification in Bengali Literature: a Comparative Analysis. *InProc. of 24th International Conference on Computational Linguistics (Coling, 2012)*, Mumbai, India, pp. 41-50.

[10] Chakraborty, T. (2012). *Multiword Expressions: Towards Identification to Applications*, LAP LAMBERT Academic Publishing GmbH & Co., KG Heinrich-Bocking-Str. 6-8, 66121, Saarbrucken, Germany, ISBN 978-3-659-24956-3.

[11] Chakraborty, T., Bandyopadhyay, S. (2010). Identification of Reduplication in Bengali Corpus and their Semantic Analysis: A Rule Based Approach. *In proc. of the 23rd International Conference on Computational Linguistics (COLING 2010), Workshop on Multiword Expressions: from Theory to Applications (MWE 2010).* Beijing, China, pp 72-75.

[12] Chakraborty, T., Das, D., Bandyopadhyay, S. (2011). Semantic Clustering: an Attempt to Extract Multiword Expressions in Bengali. *In Proc. of Multiword Expressions: from Parsing and Generation to the Real World (MWE 2011), 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies (ACL-HLT 2011)*, Portland, Oregon, USA, pp. 8-11.

[13] Chakraborty, T., Pal, S., Mondal, T., Saikh, T., Bandyopadhyay, S. (2011). Shared task system description: Measuring the Compositionality of Bigrams using Statistical Methodologies. *In Proc. of Distributional Semantics and Compositionally (DiSCo), The 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies (ACL-HLT 2011)*, Portland, Oregon, USA, pp. 38 – 42.

[14] Chattopadhyay, S. K. (1992). Bhasa-Prakash Bangala Vyakaran, Third Edition.

[15] Church, K. W., Hans, P. (1990). Word Association Norms, Mutual Information and Lexicography. *In Proc. of 27th Association for Computational Linguistics (ACL)*, vol. 16(1), pp. 22-29.

[16] Cohen, J. (1960). A coefficient of agreement for nominal scales. *Educational and Psychological Measurement*, vol. 2, pp. 37–46.

[17] Das, D., Pal, S., Mondal, T., Chakraborty, T., Bandhopadhyay, S. (2010). Automatic Extraction of Complex Predicates in Bengali. *In Proc. of Multiword Expressions: from Theory to Applications (MWE 2010), The 23rd International Conference on Computational Linguistics (COLING 2010)*, Beijing, China, pp. 37-45.

[18] Dasgupta, S., Khan, N., Sarkar, A. I., Pavel, D. S. H., Khan, M. (2005). Morphological Analysis of Inflecting Compound Words in Bengali. *In Proc. of the 8th International Conference on Computer and Information Technology (ICCIT)*, Bangladesh.

[19] Diab, M., Bhutada, P. (2009). Verb noun construction MWE token supervised classification. *In proc. of the Workshop on Multiword Expressions*, Singapore, pp. 17-22.

[20] Dias, G. H. (2003). Multiword Unit Hybrid Extraction. *In proc. of the Association for Computational Linguistics, Workshop on Multiword Expressions: Analysis, Acquisition and Treatment*, pp. 41-48.

[21] Dunning, T. (1993). Accurate Method for the Statistic of Surprise and Coincidence. *In Computational Linguistics*, pp. 61-74.

[22] Dwork, C., Kumar, R., Naor, M., Sivakumar, D. (2001). Rank aggregation methods for the web. *In proc. of Conference on the World Wide Web (WWW)-ACM*, New York, pp. 613-622.

[23] Ekbal, A., Haque, R., Bandyopadhyay, S. (2008). Maximum Entropy Based Bengali Part of Speech Tagging. *In proc. of Advances in Natural Language Processing and Applications Research in Computing Science*, pp. 67-78.

[24] Fazly, A., Stevenson, S. (2007). Distinguishing subtypes of multiword expressions using linguistically-motivated statistical measures. *In proc. of Association for Computational Linguistics, Workshop on a Broader Perspective on Multiword Expressions*. Prague, Czech Republic, pp. 9-16.

[25] Fellbaum, C. (1998). *WordNet: An Electronic lexical Database*. MIT Press, Cambridge, USA.

[26] Kilgarriff, A., and Rosenzweig, J. (2000). Framework and results for English SENSEVAL. *Computers and the Humanities. Senseval Special Issue*, 34(1-2), pp. 15-48.

[27] Korkontzelos, I, Manandhar, S. (2009). Detecting Compositionality in Multi-Word Expressions. *In proc. of the Association for Computational Linguistics-IJCNLP*, Singapore, pp. 65-68.

[28] Kunchukuttan, F. A., Damani, O. P. (2008). A System for Compound Noun Multiword Expression Extraction for Hindi. *In proc. of 6th International Conference on Natural Language Processing (ICON).* pp. 20-29.

[29] Passonneau, R. J. 2006. Measuring agreement on set-valued items (MASI) for semantic and pragmatic annotation. *Language Resources and Evaluation.*

[30] Sag, I., Baldwin, T., Bond, F., Copestake, A., Flickinger, D. (2002). Multiword Expressions: A Pain in the Neck for NLP. *In Proc. of Conference on Intelligent Text Processing and Computational Linguistics (CICLING)*, pp. 1-15.

[31] Singhal, A. (2001). Modern Information Retrieval: A Brief Overview. *Bulletin of the IEEE Computer Society Technical Committee on Data Engineering*, 24 (4), pp. 35–43.

[32] Tanaka, T., Baldwin, T. (2003). Noun-Noun Compound Machine Translation: A Feasibility Study on Shallow Processing. *In Proc. of the Association for Computational Linguistics, Workshop on Multiword Expressions: Analysis, Acquisition and Treatment*, Sapporo, Japan, pp. 17–24.

[33] Venkatpathy, S., Joshi, A. (2009). Measuring the relative compositionality of verb-noun (V-N) collocations by integrating features. *In Proc. of Human Language Technology Conference and Conference on Empirical Methods in Natural Language Processing (HLT/EMNLP), Association for Computational Linguistics*, pp. 899–906.

[34] Wu, Z., Palmar, M. (1994). Verb semantics and lexical selection. In 32nd. Annual Meeting of the Association for Computational Linguistics, pp. 133–138.

A Fast Chaos-Based Pseudo-Random Bit Generator Using Binary64 Floating-Point Arithmetic

Michael François
INSA Centre Val de Loire, Univ. Orléans, LIFO EA 4022, Bourges, France
E-mail: michael.francois@insa-cvl.fr

David Defour and Christophe Negre
Univ. Perpignan Via Domitia, DALI F-66860, LIRMM UMR 5506 F-34095, Perpignan, France
E-mail: {david.defour,christophe.negre}@univ-perp.fr

Keywords: PRBG, pseudo-random, logistic map, IEEE-754, cryptography

Received: June 27, 2013

Chaos-based cryptography is widely investigated in recent years, especially in the field of random number generators. The paper describes a novel pseudo-random bit generator (PRBG) based on chaotic logistic maps. Three logistic maps are combined in the algorithmic process, and a block of 32 random bits is produced at each iteration. The binary64 double precision format is used according to the IEEE 754-2008 standard for floating-point arithmetic. This generator provides a considerable improvement of an existing generator in the literature. Rigorous statistical analyses are carefully conducted to evaluate the quality and the robustness of the PRBG. The obtained results showed the relevance of the proposed generator, which is suitable even for real-time applications.

Povzetek: V članku je opisan hitri psevdo-naključni generator za kriptiranje.

1 Introduction

The generation of pseudo-random bits (or numbers) plays a critical role in various applications such as: statistical mechanics, numerical simulations, gaming industry, communication systems, cryptographic protocols and many others [1]. In practice, the generation of such numbers with randomness properties is an open problem and continues to be investigated. There are two main classes of generators: software and physical generators.

For the software generators, the term "pseudo-random" is applied to indicate that, the generator is defined as an algorithm allowing to produce sequences of bits with randomness properties. From a single initial seed, these generators will always produce the same outputs. The assets of such generators are: a fast execution time, repeatability and reproducibility of the pseudo-random sequences. The second class of generators exploits physical random phenomena for the generation, but is not discussed here.

Some basic techniques are often used for generating pseudo-random numbers, such as: linear recurrence [2], non-linear congruence [3], linear feedback shift register (LFSR) [4], cellular automata [5], discrete logarithm problem [6], quadratic residuosity problem [7], etc. Generally, the security of a cryptographic generator is based on the difficulty to solve the related mathematical problem. Beyond the security, such kind of generator is sometimes too slow due to heavy computational instructions. For example, the Blum Blum Shub algorithm [7] has a security proof, assuming the computational difficulty of the quadratic residuosity problem. The algorithm is also proven to be secure, relatively to the difficulty of integer factorization problem. However, the generator is impractical unless extreme security is needed. The Blum-Micali algorithm [6] presents also an unconditional security proof based on the difficulty of the discrete logarithm problem, but is also ineffective.

One interesting way to design pseudo-random generators can be found in chaos theory [8, 9, 10]. Indeed, chaotic systems are characterized by their high sensitivity to initial parameters and some properties like ergodicity, mixing property and high complexity [8, 11]. A secret parameter should be sensitive enough to ensure the so-called avalanche property. A small deviation in the initial conditions should cause a large modification in the output, that makes chaotic systems very attractive for pseudo-random number generation. These generators commonly use chaotic logistic maps and produce large pseudo-random sequences. For a high security level, it is necessary to combine several logistics maps, in order to increase the complexity of the cryptosystem. But, this is not always sufficient, because a rigorous analysis is more appropriate to evaluate the randomness level and the global security of the generator.

In this paper, a new PRBG combining three chaotic logistic maps is presented. It provides a significant improvement on security and performance, of the generator proposed by Patidar et al. [12]. The proposed algorithm uses the binary64 floating-point arithmetic and produces at each iteration a block of 32 random bits. The pseudo-random sequences passed successfully the various statistical tests related to the randomness and correlation. The assets of the PRBG are: high sensitivity to initial seeds, high level of randomness and fast execution time. The paper is structured as follows, in Sec. 2 the used chaotic logistic map and the description of the Patidar's algorithm are given. Section 3, presents a detailed description of our algorithm and a brief discussion about the floating-point representation. The statistical analysis is given in Sec. 4. The security aspect of the PRBG is discussed in Sec. 5, before concluding.

2 Background

2.1 The chaotic logistic map

Frequently used in chaos theory as well as in chaos-based cryptosystems, the form of the logistic map is given by:

$$F(X) = \beta X(1 - X), \tag{1}$$

with β between 3.57 and 4.0 [13]. Its chaotic behavior has been widely studied and several generators have already used such logistic map for generating pseudo-random numbers [14, 15, 16, 17]. To avoid non-chaotic behaviour (island of stability, oscillations, ...), the value of β should be near 4.0, which corresponds to a highly chaotic behaviour. The logistic map is used under the iterative form:

$$X_{n+1} = \beta X_n(1 - X_n), \forall n \geq 0, \tag{2}$$

where the starting seed X_0 is a real number belonging to the interval $]0, 1[$. All the computed elements X_n are also real numbers in $]0, 1[$.

2.2 About the algorithm of Patidar

Patidar et al. [12] have proposed a PRBG based on two logistic maps. The algorithm starts from random independent initial seeds X_0, Y_0, belonging to $]0, 1[$. The chosen value of β is 4 and the two logistic maps are given by:

$$\begin{aligned} X_{n+1} &= 4X_n(1 - X_n), \forall n \geq 0, & (3)\\ Y_{n+1} &= 4Y_n(1 - Y_n), \forall n \geq 0. & (4) \end{aligned}$$

The main idea of the algorithm is very simple and consists to compare the outputs of both the logistic maps in the following way:

$$g(X_{n+1}, Y_{n+1}) = \begin{cases} 1 & if\, X_{n+1} > Y_{n+1} \\ 0 & if\, X_{n+1} \leq Y_{n+1} \end{cases}$$

Even if the idea is interesting, the algorithm presents several weaknesses:

1. Only one bit is generated after each iteration, that corresponds to a very low throughput according to the relevance of the logistic maps.
2. The sequences produced with nearby seed values are extremely correlated.
3. The seed space has a much lower entropy than 128, due to the existing correlation between the pseudo-random sequences. Therefore, the generator presents weak or degenerate seeds.
4. At a given iteration n, in the case of eventual collision between X_{n+1} and Y_{n+1} (which is possible), the output bit will always be 0 until the end of the output sequence.

The algorithm proposed in this paper also combines several chaotic logistic maps, but is designed to avoid all those weaknesses and then ensure a better security.

3 The proposed PRBG

3.1 Floating-point representation

As we know, digital computers use binary digits to represent numbers. In the case of real numbers, there are two representation formats: fixed-point and floating-point formats. To represent integers or real numbers with a fixed precision, it is more suitable to adopt the first format. The second format can support a much wider range of values. Nowadays, the floating point arithmetic is standardized by IEEE/ANSI [18]. Two different floating-point formats are defined: single precision (binary32) and double precision (binary64). In this paper, we only focus on binary64 floating-point format, which is generally used to achieve a higher simulation precision for the study of chaotic systems.

Binary64 has two infinities, two kinds of NaN (i.e. Not a Number) and the set of finite numbers. Each finite number is described by three fields: s a sign represented on one bit (1 indicating negative), e a biased exponent represented on 11 bits and m a mantissa represented on 52 bits (see Figure 1). The bits of the mantissa can be divided into two blocks of 20 bits and 32 bits and the treatment is applied on the block of 32 bits (mantissa1).

3.2 Description of the algorithm

As in the paper of Patidar et al., our algorithm uses the same type of chaotic logistic map given by Eq. 1. In our case, the value of β is fixed to 3.9999 that corresponds to a highly chaotic case [19, 20]. Indeed, the Lyapunov exponent [21, 22] measures the chaotic behavior of a function and the corresponding Lyapunov exponent of the logistic map for $\beta = 3.9999$ is 0.69 very close to its maximum which is 0.59. The chaotic logistic map is used under the iterative form:

$$X_{n+1} = 3.9999X_n(1 - X_n), \forall n \geq 0, \tag{5}$$

where the starting seed X_0 is a real number that belongs to $]0, 1[$. All the computed elements X_n are also real numbers

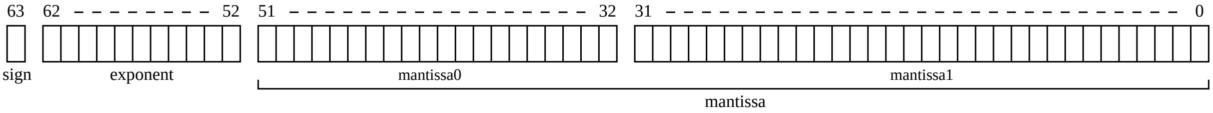


Figure 1: Floating-point representation in binary64 format.

in $]0,1[$.

Our algorithm takes into account the various weaknesses of the algorithm proposed by Patidar et al. Thus, to have a large space of output sequences, three logistic maps are used during the generation process. The same value of β is used for each one and the corresponding equations are:

$$X_{n+1} = 3.9999X_n(1-X_n), \forall n \geq 0, \quad (6)$$

$$Y_{n+1} = 3.9999Y_n(1-Y_n), \forall n \geq 0, \quad (7)$$

$$Z_{n+1} = 3.9999Z_n(1-Z_n), \forall n \geq 0. \quad (8)$$

For each computed value X_n, Y_n and Z_n, a binary64 floating-point representation is used as shown in Figure 1. The algorithmic principle is simple and consists at each iteration, to apply a xor operation on the 32 bits of mantissa1 of the three output elements X_n, Y_n and Z_n. Thus, the algorithm allows to produce 32 random bits per iteration and therefore increase the throughput of the generator. The operating principle of the algorithm is shown in Figure 2. As one can see, the seeds from which the generation process starts are X_k, Y_k and Z_k. Indeed, for nearby seed values, the elements X_n, Y_n and Z_n are almost identical in the first rounds. Thus, to completely decorrelate the beginning of the pseudo-random sequences, it is necessary to start the generation only at the kth iteration. The number of preliminary rounds k and the way to choose the initial seeds are presented in Sec. 3.3. The implementation of the algorithm in C language is simple: just include the file $ieee754.h$ and use the defined functions for extracting the bits of mantissa1 for each computed element X_n, Y_n and Z_n.

3.3 The choice of initial parameters

3.3.1 Initial seed selection

To improve the randomness quality of the generated sequences, the choice of the initial seed values should not be neglected. The coefficient values of the elements X_n, Y_n and Z_n, belong to $]0,1[$. Due to symmetric structure of the logistic map, it is necessary to choose the starting seeds in one of the two half-intervals (here $]0,2^{-1}[$) to avoid similar trajectories. In binary64 floating-point format, the computed term $(1-X)$ is equal to 1.0 for any X in $]0,2^{-53}[$, then for a seed value in $]0,2^{-53}[$, the computed value of Eq. 2 is equivalent to βX_n. To avoid such problem, initial seed values must be chosen in $]2^{-53},2^{-1}[$.
The three initial seeds must be different, then the difference $\delta_{[2]}$ between the values should be representable in binary64. The value of $\delta_{[2]}$ is in the worst case 2^{-53}, which corresponds to $\log_{10}(2^{53})$ (≈ 15.955) decimal digits. To have a significant difference we choose $\delta_{[10]} = 10^{-15}$, which corresponds to $\delta_{[2]} = 2^{-49.8289}$. Thus, to avoid identical trajectories, the difference between each initial seed should be at least $\delta_{[2]} = 2^{-49.8289}$.

3.3.2 Number of preliminary rounds

In the case where the values of initial seeds (X_0, Y_0 and Z_0) are very close, the beginnings of chaotic trajectories are almost similar. To avoid such problem, it is necessary to apply some preliminary rounds before starting to produce the random bits. Thus, it is necessary to see at which number of iterations, the difference $\delta_{[2]}$ begins to be propagated. We consider that the initial seed is $X_0 = \delta_{[2]} = 2^{-49.8289}$, and the obtained trajectory with the Eq. 5 is shown in Figure 3.

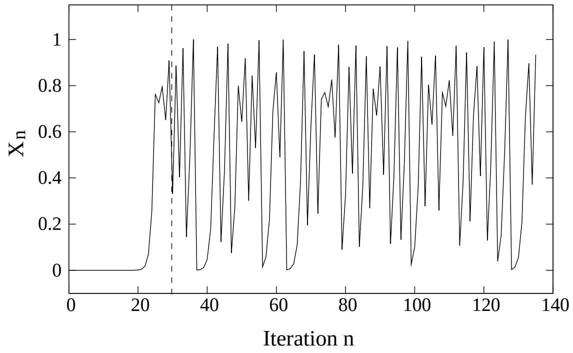


Figure 3: Trajectory of the chaotic logistic map given in Eq. 5, for $n = 135$ and $X_0 = 2^{-49.8289}$.

One can see that, the trajectory starts to oscillate almost from the 30th iteration. Thus, the generation of random bits will begin from the iteration 30. That allows to decorrelate the outputs of the PRBG, and then increase the sensitivity related to the initial seeds.

4 Statistical analysis

The quality of the output sequences produced by any PRBG is the crucial element. Indeed, the sequences should present individually a high level of randomness and be decorrelated with each other, whatever the initial seed values. Therefore, a statistical analysis should be carefully conducted to prove the quality of the pseudo-random sequences.

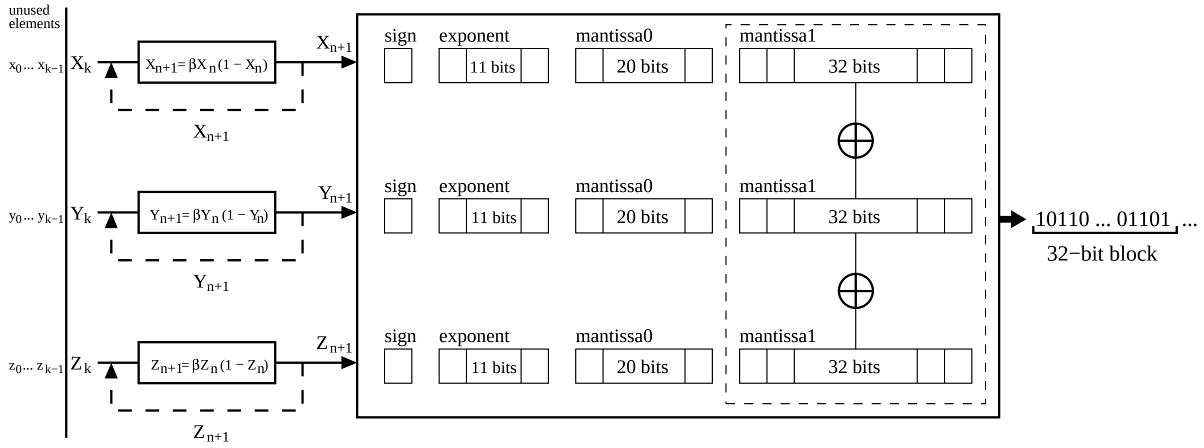


Figure 2: The operating principle of the proposed PRBG.

4.1 Randomness evaluation

The analysis consists in evaluating the randomness level of the sequences generated by the PRBG. In the literature, various statistical tests exist for analysing the randomness level of sequences. The NIST (National Institute of Standards and Technology of the U.S. Government) proposes a battery of tests that can be applied on the binary sequences [23]. One can also find other known libraries such as TestU01 [24] or the DieHARD suites [25]. Here, the sequences are evaluated through statistical tests suite NIST. Such suite consists in a statistical package of fifteen tests developed to quantify and to assess the randomness of binary sequences, produced by pseudo-random number generators. Here, we define three approaches for testing the randomness level of sequences. Let N be the total number of generated sequences and the binary size of each sequence is $M = 32 \times B$, with B the number of 32-bit blocs. The three approaches are:

1. **APP-1** (individual sequences): the produced sequences are individually tested and the results are given as ratio of success relatively to a threshold determined from the total number (N) of tested sequences. Such approach indicates the global randomness level of the tested sequences.

2. **APP-2** (concatenated sequence): all the individual sequences are concatenated to form a new single sequence. The randomness level of the constructed sequence is analysed through the NIST tests. The constructed sequence should pass the tests whether the original sequences are truly decorrelated and random.

3. **APP-3** (resulting sequences): all the sequences are superimposed on each other (forming a matrix), and new sequences are constructed from columns. Thus, B resulting sequences of binary size $32 \times N$ are constructed, by collecting for each position $1 \leq j \leq B$, the 32-bit bloc of each sequence. If the original sequences are really random, the resulting sequences should also be random (with B as large as N) and then pass the NIST tests. Such approach is very interesting, in the case of generating sequences by nearby seed values, and allows to detect some hidden linear structures between the original sequences.

These approaches are used to analyse a subset of generated sequences. In the case of very distant initial seed values, the corresponding chaotic trajectories are different, and allow to produce good pseudo-random sequences. The worst case occurs when closed seed values are used, because that can lead to highly correlated output sequences. That is why, the analysis is achieved on sequences generated from nearby initial seed values. Here, a subset of $N = 16000$ pseudo-random sequences is produced, where the binary size of each sequence is 32000 (i.e. $B = 1000$). We choose arbitrarily, one starting seed value $X_0 = 0.24834264038461704925$, and then $Y_0 = X_0 + \delta_{[2]}$ and $Z_0 = Y_0 + \delta_{[2]}$, with $\delta_{[2]} = 2^{-49.8289}$. These three seeds allow to generate one pseudo-random sequence. The other sequences, are generated from X_0, Y_0 and by incrementing of $\delta_{[2]}$ the last seed value Z_0 in a simple loop.
In the case of Patidar's algorithm, only two initial seeds are needed to produce a pseudo-random sequence. Here, the first two seeds values are given by: $X'_0 = Y_0$ and $Y'_0 = Z_0$. For generating the other sequences, the same strategy is applied and consists to make a loop by incrementing of $\delta_{[2]}$ the seed Y'_0. It should be noted that, for a better comparison, the same coefficient β (i.e. 3.9999) is used for the logistic map in the Patidar's algorithm.
The results of NIST tests obtained for the two algorithms are presented in Table 4.1 and Table 4.1. For approach **APP-1** (resp. **APP-3**), the acceptable proportion should lie above 98.76% (resp. 98.00%) and does not concern the tests *Random Excursions-(Variant)*. For **APP-2**, a sequence passes a statistical test for $p_{value} \geq 0.01$ and fails otherwise. For the tests *Non-Overlapping* and *Random Excursions-(Variant)*, only the smallest percentage of all sub-tests is given. For individual sequences, the *Universal*

test is not applicable due to the size of initial sequences. One can remark that, for the proposed PRBG, all the tested sequences pass successfully the NIST tests. These results show clearly the quality of the tested sequences. For Patidar's algorithm, individually, the sequences are not enough random, because there are many tests that are not successful, for example: *Runs, Overlapping* or even *Serial* tests. The results of approaches **APP-2** and **APP-3** show that, the tested sequences are extremely correlated. One should know that, in the article of Patidar et al., each sequence is produced from a randomly chosen initial seed belonging to $]0,1[$, then the seeds were too different from each other. That is why this problem has not been detected.

4.2 Correlation evaluation

A part of the correlation evaluation has already been done by applying the NIST tests (**APP-2** and **APP-3**). Here, two additional methods are used to analyse the correlation between the pseudo-random sequences. Firstly, the correlation between sequences is evaluated globally by computing the Pearson's correlation coefficient [26] and secondly, by using the Hamming distance.

4.2.1 Pearson's correlation coefficient

The analyse consists to compute the Pearson's correlation coefficient between each pair of sequences, and to present the distribution of the values through a histogram. Consider a pair of sequences such as: $S_1 = [x_0, \dots, x_{B-1}]$ and $S_2 = [y_0, \dots, y_{B-1}]$. Therefore, the corresponding correlation coefficient is:

$$C_{S_1,S_2} = \frac{\sum\limits_{i=0}^{B-1} (x_i - \overline{x}) \cdot (y_i - \overline{y})}{\left[\sum\limits_{i=0}^{B-1} (x_i - \overline{x})^2\right]^{1/2} \cdot \left[\sum\limits_{i=0}^{B-1} (y_i - \overline{y})^2\right]^{1/2}}, \quad (9)$$

where x_i and y_i are 32-bit integers, $\overline{x} = \sum\limits_{i=0}^{B-1} x_i/B$ and $\overline{y} = \sum\limits_{i=0}^{B-1} y_i/B$, the mean values of S_1 and S_2, respectively. Two uncorrelated sequences are characterized by $C_{S_1,S_2} = 0$. The closer the value of C_{S_1,S_2} is to ± 1, the stronger the correlation between the two sequences. In the case of two independent sequences, the value of C_{S_1,S_2} is equal to 0. Here we use the same subsets of 16000 sequences, and the coefficients C_{S_1,S_2} are computed. For the two algorithms, the histograms are shown in Figure 4. For the proposed PRBG, around 99.56% of the coefficients have an absolute value smaller than 0.09, then only a small correlation is detected. In the case of Patidar's PRBG, around 99.26% of the coefficients have an absolute value greater than 0.33, that means the sequences are highly correlated.

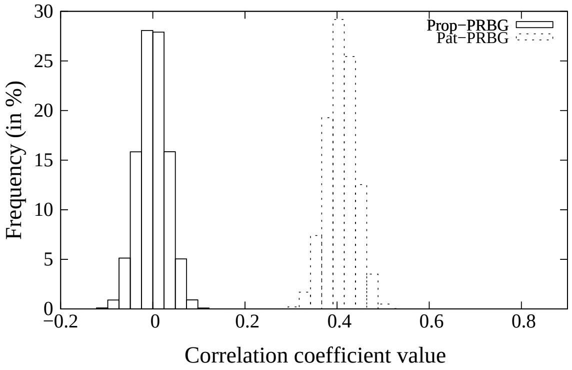


Figure 4: Histogram of Pearson's correlation coefficient values on interval $[-0.1, 0.1]$ (resp. $[-0.5, 0.5]$), for the proposed (resp. Patidar's) PRBG.

4.2.2 Hamming distance

Another type of correlation based on the bits of produced pseudo-random sequences is analysed. Given two binary sequences $S = [s_0, \dots, s_{M-1}]$ and $S' = [s'_0, \dots, s'_{M-1}]$ of same length (M), the Hamming distance is the number of positions where they differ. The distance is given as:

$$d(S, S') = \sum_{j=0}^{M-1} (s_j \oplus s'_j). \quad (10)$$

For truly random binary sequences, the value of $d(S, S')$ should be around $M/2$, that corresponds to the proportion 0.50. This distance is computed between each pair of generated sequences ($N = 16000$), and all values are represented through a histogram. For the two algorithms, the histograms are shown in Figure 5. One can see that for our algorithm, all the proportions of computed Hamming distances are around the mid-value 0.50 and almost 99.95% of the coefficients belong to $]0.49, 0.51[$. In the second case, the values are around 0.28, and near 99.83% of the coefficients belong to $]0.26, 0.30[$. The results show that, the

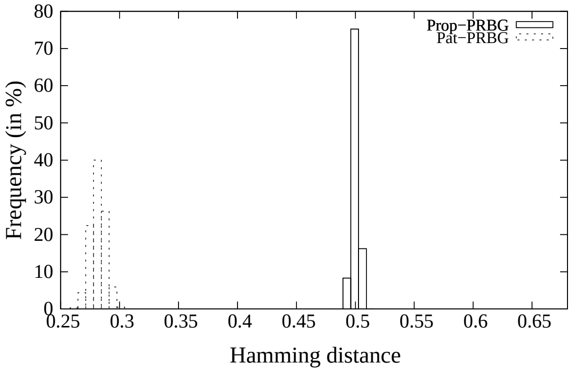


Figure 5: Histogram of Hamming distances on interval $[0.48, 0.52]$ (resp. $[0.25, 0.30]$), computed between each pair of sequences for the proposed (resp. Patidar's) PRBG.

Test Name	Proposed PRBG					
	APP-1		**APP-2**		**APP-3**	
	r_1 (in %)	Result	p_{value}	Result	r_2 (in %)	Result
Frequency	99.16	Success	0.378240	Success	99.50	Success
Block-Frequency	99.10	Success	0.905858	Success	98.80	Success
Cumulative Sums (1)	99.17	Success	0.447272	Success	99.40	Success
Cumulative Sums (2)	99.12	Success	0.259837	Success	99.30	Success
Runs	98.90	Success	0.654035	Success	99.20	Success
Longest Run	98.90	Success	0.717020	Success	98.70	Success
Rank	98.86	Success	0.239335	Success	99.00	Success
FFT	98.76	Success	0.485387	Success	99.00	Success
Non-Overlapping	99.30	Success	0.012842	Success	98.20	Success
Overlapping	99.00	Success	0.935098	Success	98.80	Success
Universal	-	-	0.196700	Success	98.60	Success
Approximate Entropy	98.91	Success	0.199988	Success	99.00	Success
Random Excursions	97.56	Success	0.012412	Success	98.60	Success
Random Ex-Variant	97.56	Success	0.024851	Success	97.62	Success
Serial (1)	98.92	Success	0.379823	Success	99.30	Success
Serial (2)	99.05	Success	0.856303	Success	99.20	Success
Linear Complexity	98.84	Success	0.098641	Success	99.00	Success

Table 1: The results of NIST tests for the proposed PRBG on the 16000 sequences. The ratio r_1 (resp. r_2) of p_{value} passing the tests are given for **APP-1** (resp. **APP-3**). For the approach **APP-2** the corresponding p_{value} is given.

Test Name	Patidar's PRBG					
	APP-1		**APP-2**		**APP-3**	
	r_1 (in %)	Result	p_{value}	Result	r_2 (in %)	Result
Frequency	99.84	Success	0.000000	Failure	02.80	Failure
Block-Frequency	99.99	Success	0.989313	Success	16.00	Failure
Cumulative Sums (1)	99.80	Success	0.000000	Failure	03.10	Failure
Cumulative Sums (2)	99.74	Success	0.000000	Failure	03.00	Failure
Runs	29.25	Failure	-	-	00.60	Failure
Longest Run	00.00	Failure	0.000000	Failure	00.00	Failure
Rank	98.83	Success	0.442618	Success	04.00	Failure
FFT	98.70	Success	0.000000	Failure	00.00	Failure
Non-Overlapping	75.96	Failure	0.000000	Failure	71.80	Failure
Overlapping	00.00	Failure	0.000000	Failure	99.00	Success
Universal	-	-	0.000000	Failure	00.40	Failure
Approximate Entropy	00.00	Failure	0.000000	Failure	00.00	Failure
Random Excursions	98.55	Success	-	-	50.00	Failure
Random Ex-Variant	97.82	Success	-	-	93.75	Success
Serial (1)	95.10	Failure	0.000000	Failure	00.00	Failure
Serial (2)	98.99	Success	0.000000	Failure	00.00	Failure
Linear Complexity	98.81	Success	0.283356	Success	99.00	Success

Table 2: The results of NIST tests for the PRBG of Patidar et al., on the 16000 sequences. The ratio r_1 (resp. r_2) of p_{value} passing the tests are presented for the approach **APP-1** (resp. **APP-3**). For **APP-2** the corresponding p_{value} is given.

sequences are correlated for the Patidar's PRBG, then the algorithm is not enough sensitive to initial seed values. For more sensitivity, one must choose very different seed values, which reduces considerably the key space and then the security of the PRBG.

4.3 Seed sensitivity

A small deviation from the initial seeds, should cause a large variation in the output sequences. Actually, in the NIST tests (**APP-2** and **APP-3**, Sec. 4.1) and the correlation evaluation (Sec. 4.2), the sensitivity related to the seeds was indirectly analysed. To make an additional analysis, a large size of pseudo-random sequences is considered. Here, the number of 32-bit blocs is $B = 10000000$, then the binary size $M = 320000000$. A pseudo-random sequence ($Seq1$) is produced using the seed values: $X_0 = 0.32164872553014364784, Y_0 = X_0 + \delta_{[2]}$ and $Z_0 = Y_0 + \delta_{[2]}$. Two others sequences ($Seq2$ and $Seq3$) are produced by adding the value of $\delta_{[2]}$ on the last seed value Z_0. Between each pair of sequences, the correlation analysis is done by computing the linear correlation coefficient of Pearson, the correlation coefficient of Kendall [27] and the Hamming distance. The same analysis is applied on the Patidar's algorithm with the starting seeds $X'_0 = Y_0$ and $Y'_0 = Z_0$. The results are given in Table 4.3 and show that: our algorithm is highly sensitive to initial seeds, whereas in the case of the Patidar's algorithm, the sensitivity is extremely weak.

4.4 Speed analysis

Another important aspect for any PRBG is the execution time of the algorithm. Indeed, in real-time applications, the temporal constraint about the performance of a process is as considerable as the final results of the process. The speed evaluation is achieved on a work computer with processor: Intel(R) Xeon(R) CPU E5410 @ 2.33 GHz × 4. The source code is compiled using GCC 4.6.3 on Ubuntu (64 bits). The results are presented in Table 4.4 and one can see that, with no optimization option (-O0), the proposed algorithm enables to produce around 2.62 Gbits per second. However, with the classical optimization option (-O1), the throughput is approximately 80 Gbits per second.

PRBG	Speed (Gbits/s)	
	-O0	-O1
Proposed	2.62	80.00
Patidar's	0.06	1.18

Table 4: Comparison of speed between the two algorithms by using the options "-O0" and "-O1".

The Table 4.4 presents the approximative throughput of some known pseudo-random number generators. One can remark that, the throughput of our generator is almost in the same order than CURAND.

Generator	Speed (Gbits/s)	
	GT120 GPU (4 cores)	GTX260 GPU (27 cores)
MTGP11213	41.88	340.42
CURAND	96.97	533.33

Table 5: The approximative throughput in Gbits/s for MTGP11213 (*Mersenne Twister for Graphic Processor*) and CURAND (*NVIDIA CUDA Random Number Generation library*).

5 Security analysis

Some points related to the security of the PRBG are discussed here, such as: the size of the seed space, the period length of the logistic map and some basic-known attacks (brute-force attack and differential attack).

5.1 Seed space

Given today's computational resources, a seed space of size smaller than 2^{128} is not secure enough. A robust PRBG should have a large key space, to allow a large choice for the pseudo-random number generation. In order to enlarge the key space, three chaotic logistic maps are used during the generation process. Each logistic map needs to be initialized with a seed corresponding to a binary64 floating-point number, selected from $]2^{-53}, 2^{-1}[$. Knowing that the difference between each seed value is $2^{-49.8289}$, this allows to have $2^{48.8289}$ possible choices of initial seeds. Thus, the total number of choices for the three initial seeds is:

$$2^{48.8289} \times [2^{48.8289} - 1] \times [2^{48.8289} - 2],$$

or about $2^{146.50}$. In the case of Patidar's algorithm, the entropy of the seed space is much smaller than 128. Indeed, the algorithm uses only two logistic maps and possesses a weak sensitivity, that requires to choose very distant seeds to produce secure outputs.

5.2 Period length of the logistic map

The period length is a fundamental indicator of any PRBG. A generator should have a reasonably long period before its output sequence repeats itself, for avoiding attacks. The length of the period will indicate the maximal secure size for the producible pseudo-random sequences. The idea is to determine the cycle formed by each chaotic trajectory according to the different starting seed values. In a period-p cycle, $X_k = F^p(X_k)$ for some X_k, where F^p is the pth iterate of F, *e.g.*, $F^3(X) = F(F(F(X)))$ for $p = 3$.

The GNU MPFR library [29] is used to vary the bits of the mantissa for analysing the cycles of the logistic map. The Figure 6 shows the length of cycles, when the bits of mantissa vary between 10 and 25. In this case, the logistic map has very small cycle lengths. The Figure 7 indicates at the the same time the corresponding occurrences.

PRBG	Tests	$Seq1/Seq2$	$Seq1/Seq3$	$Seq2/Seq3$
Proposed algorithm	Pearson Corr. Coef.	-0.000460	0.000210	-0.000536
	Kendall Corr. Coef.	-0.000127	-0.000377	-0.000201
	Hamming Distance	0.499985	0.499986	0.500010
Patidar's algorithm	Pearson Corr. Coef.	0.328897	0.328990	0.328856
	Kendall Corr. Coef.	0.245210	0.242712	0.242441
	Hamming Distance	0.333467	0.333379	0.333313

Table 3: Comparison using the Pearson's and Kendall's correlation coefficients, and also Hamming distance (in proportion) between each pair of sequences (Seq_1, Seq_2 and Seq_3), produced from slightly different seeds.

In binary32 format, the obtained smallest (resp. longest) cycle length is equal to 1 (resp. 3055). Also, the logistic map has numerous pathological seeds (corresponding in minimum cycles of length 1) and globally the cycle lengths are too small. Therefore, the binary32 format is not appropriate and must be avoided, when implementing a PRBG with such logistic map. Besides, this result is consistent with that published by Persohn and Povinelli [30].

For binary64 format, the cycle lengths are much longer. Due to the large size of the binary format, it is difficult to analyse all the corresponding trajectories. Only a reasonable set of randomly chosen seeds is considered. The length of the smallest cycle is 2169558 ($\approx 2^{21.04}$), while for the longest cycle is 40037583 ($\approx 2^{25.25}$). Here, the computed cycle lengths are in the same order as those given in [28], and no pathological seed was found. This format was not studied by Persohn and Povinelli, and it is a format that benefits to the logistic map. Also, the used parameter λ (equal to 4) does not provide the best chaotic behavior. In our case, combining three chaotic logistic maps allows to increase the length of the global resulting cycle, which is given by the LCM of the three cycle lengths. Also, the value of β (here 3.9999) plays a crucial role, because it provides a better chaotic behavior of the logistic map. However, for a maximum security level, it might be better to limit the length of sequences to the smallest cycle length. The best way to avoid the problem of short period and use efficiently this PRBG, is to generate pseudo-random bit sequences of only small sizes. However in case of need, long sequences can be constructed by concatenating several ones.

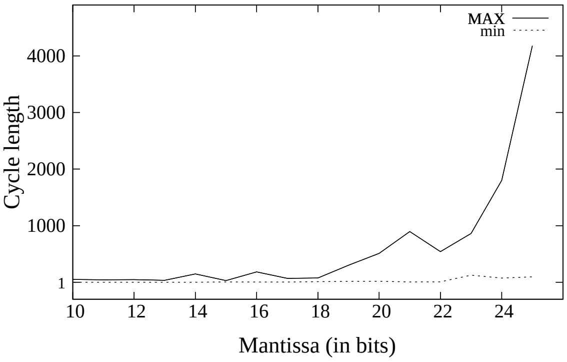


Figure 6: The curve "MAX" (resp. "min") shows the length of longest (resp. smallest) cycles, when varying mantissa bits between 10 and 25.

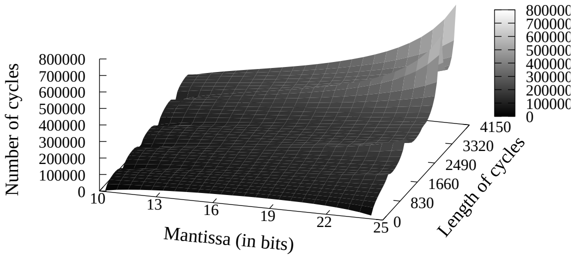


Figure 7: Representation of the length of cycles and their total numbers, when the bits of the mantissa vary between 10 and 25.

5.3 Brute-force attack

In theory, a brute-force attack [8] is an attack that can be used against any kind of PRBG. Such attack is usually utilized, when it is not easy (or possible) to detect any weakness in the algorithm, that would make the task easier. The strategy of the attack is simple and consists to check systematically all possible keys until the original key is found. On average, just half of the size of key space needs to be tested to find the initial seeds. A large key space allows to frustrate such kind of attack. Nowadays a key space of size larger than 2^{128} is computationally secure enough to resist to such attack. The proposed PRBG has a key space of size $2^{146.50}$ and we consider that, such attack can not succeed on the generator. In the case of Patidar's PRBG, the entropy of the key space is logically much smaller than 128. Thus, such attack can be envisaged for breaking the generator.

5.4 Differential attack

As a chosen-plaintext attack, the principle of such technique of cryptanalysis is to analyse the effect of a small difference in input pairs (i.e. seeds), on the difference of corresponding output pairs (i.e. sequences) [31]. This strategy allows to get the most probable key, that was used to generate the pseudo-random sequence. The initial difference may be in the form of a subtraction modulus or a xor difference and the diffusion aspect is measured by a differential probability. The proposed algorithm is designed to avoid such attack. Indeed, the initial seeds are chosen in the interval $]2^{-53}, 2^{-1}[$ and the bit-generation starts only at the 30th iteration. The results of the statistical analysis (Sec. 4) showed also that, even with a small difference on the seeds, the pseudo-random sequences are highly decorrelated from each other. Thus, we consider that the proposed PRBG should resist to the differential cryptanalysis. On the other side, the attack can be possible on the Patidar's PRBG, because the algorithm is not sensitive enough to the initial seed values.

6 Conclusion

A chaos-based PRBG, combining three chaotic logistic maps under binary64 floating-point arithmetic was presented. It provides significant improvements of an existing generator. The principle consists at each iteration, to apply a bitwise xor operator on the 32 least significant bits of mantissa, from the computed elements of logistic maps. The algorithm is fast and allows to produce pseudo-random sequences formed of 32-bit blocks. The assets of the PRBG are: the simplicity of implementation, a high randomness level for outputs, a high sensitivity related to the initial seeds and a fast execution time, allowing to use the algorithm even in real-time applications.

References

[1] F. Sun and S. Liu (2009). Cryptographic pseudo-random sequence from the spatial chaotic map. *Chaos Solitons & Fractals*, Vol. 41, No. 5, pp. 2216–2219.

[2] M. Matsumoto and T. Nishimura (1986). Mersenne twister: a 623-dimensionally equidistributed uniform pseudo-random number generator. *ACM Transactions on Modeling and Computer Simulation (TOMACS)*, Vol. 8, No. 1, pp. 3–30.

[3] J. Eichenauer and J. Lehn (1986). A non-linear congruential pseudo random number generator. *Statistische Hefte*, Vol. 27, No. 1, pp. 315–326.

[4] A.K. Varshney, S.K. Sharma and R. Singh (2012). A study on the effect of shifting on LFSR PRNG. *International Journal of Engineering*, Vol. 1, No. 5, pp. 1–7.

[5] M. Tomassini, M. Sipper, M. Zolla and M. Perrenoud (1999). Generating high-quality random numbers in parallel by cellular automata. *Future Generation Computer Systems*, Vol. 16, No. 2, pp. 291–305.

[6] M. Blum and S. Micali (1984). How to generate cryptographically strong sequences of pseudorandom bits. *SIAM journal on Computing*, Vol. 13, No. 4, pp. 850–864.

[7] L. Blum, M. Blum and M. Shub (1986). A simple unpredictable pseudo-random number generator. *SIAM journal on Computing*, Vol. 15, No. 2, pp. 364–383.

[8] G. Álvarez and S. Li (2006). Some Basic Cryptographic Requirements for Chaos-Based Cryptosystems. *International Journal of Bifurcation and Chaos*, Vol. 16, No. 8, pp. 2129–2151.

[9] H.P. Hu, L.F. Liu and N.D. Ding (2012). Pseudorandom sequence generator based on Chen chaotic system. *Computer Physics Communications*, Vol. 184, No. 3, pp. 765–768.

[10] M. François, T. Grosges, D. Barchiesi and R. Erra (2013). A new pseudo-random number generator based on two chaotic maps. *Informatica*, Vol. 24, No. 2, pp. 181–197.

[11] J.M. Bahi, C. Guyeux and Q. Wang. A pseudo random numbers generator based on chaotic iterations. application to watermarking. *In WISM 2010, International Conference on Web Information Systems and Mining*, Vol. 6318 of LNCS, pp. 202–211, Sanya, China, October 2010.

[12] V. Patidar, K.K. Sud and N.K. Pareek (2009). A Pseudo Random Bit Generator Based on Chaotic Logistic Map and its Statistical Testing. *Informatica*, Vol. 33, No. 4, pp. 441–452.

[13] R. Bose and A. Banerjee. Implementing Symmetric Cryptography Using Chaos Functions, *in: Proceedings of the 7th International Conference on Advanced Computing and Communications*, pp. 318–321, 1999.

[14] M.S. Baptista (1998). Cryptography with chaos. *Physics Letters A*, Vol. 240, No. 1, pp. 50–54.

[15] S. Cecen, R.M. Demirer and C. Bayrak (2009). A new hybrid nonlinear congruential number generator based on higher functional power of logistic maps. *Chaos Solitons & Fractals*, Vol. 42, No. 2, pp. 847–853.

[16] S. Xuan, G. Zhang and Y. Liao. Chaos-based true random number generator using image. *in: IEEE International Conference on Computer Science and Service System (CSSS), Nanjing, China*, pp. 2145–2147, 2011.

[17] M. François, T. Grosges, D. Barchiesi and R. Erra (2014). Pseudo-random number generator based on mixing of three chaotic maps. *Communications in Nonlinear Science and Numerical Simulation*, Vol. 19, No. 4, pp. 887–895.

[18] IEEE Standard for Floating-Point Arithmetic. *IEEE Std 754-2008*, pp. 1–58.

[19] N.K. Pareek, V. Patidar and K.K. Sud (2006). Image encryption using chaotic logistic map. *Image and Vision Computing*, Vol. 24, No. 9, pp. 926–934.

[20] M. François, T. Grosges, D. Barchiesi and R. Erra (2012). Image Encryption Algorithm Based on a Chaotic Iterative Process. *Applied Mathematics*, Vol. 3, No. 12, pp. 1910–1920.

[21] A. Wolf, J.B. Swift, H.L. Swinney, J.A. Vastano (1985). Determining Lyapunov exponents from a time series. *Physica D: Nonlinear Phenomena*, Vol. 16, No. 3, pp. 285–317.

[22] E. Aurell, G. Boffetta, A. Crisanti, G. Paladin, A. Vulpiani (1997). Predictability in the large: an extension of the concept of Lyapunov exponent. *Journal of Physics A: Mathematical and General*, Vol. 30, No. 1, pp. 1–26.

[23] A. Rukhin, J. Soto, J. Nechvatal, M. Smid, E. Barker, S. Leigh, M. Levenson, M. Vangel, D. Banks, A. Heckert, J. Dray and S. Vo. A Statistical Test Suite for Random and Pseudorandom Number Generators for Cryptographic Applications. *NIST Special Publication Revision 1a 2010.*

[24] P. L'ecuyer and R. Simard (2007). TestU01: A C library for empirical testing of random number generators. *ACM Transactions on Mathematical Software*, Vol. 33, No. 4, Article 22.

[25] G. Marsaglia (1996). Diehard: a battery of tests of randomness. http://stat.fsu.edu/geo/diehard.html

[26] J. Rodgers and W. Nicewander (1988). Thirteen ways to look at the correlation coefficient. *The American Statistician*, Vol. 42, No. 1, pp. 59–66.

[27] M.G. Kendall. Rank correlation methods. *Fourth ed., Griffin, London*, 1970.

[28] J. Keller and H. Hanno. Period lengths of chaotic pseudo-random number generators, *in: Proceedings of the Fourth IASTED International Conference on Communication, Network and Information Security*, pp. 7–11, 2007.

[29] The GNU MPFR library. http://www.mpfr.org

[30] K.J. Persohn and R.J. Povinelli (2012). Analyzing logistic map pseudorandom number generators for periodicity induced by finite precision floating-point representation. *Chaos Solitons & Fractals*, Vol. 45, No. 3, pp. 238–245.

[31] E. Biham and A. Shamir. Differential Cryptanalysis of the Data Encryption Standard. *Springer-Verlag*, 1993.

Optimizing the Classification Cost using SVMs with a Double Hinge Loss

Amirou Ahmed†, Ould-Abdeslam Djaffar‡ and Zidelmal Zahia†
†Mouloud Mammeri University , Tizi-Ouzou, Algeria
E-mail: a-amirou@mail.ummto.dz, djaffar.ould-abdeslam@uha.fr

Aidene Mohamed† and Merckle Jean‡
‡MIPS Laboratoiry, Haute Alsace University, France
E-mail: m-aidene@mail.ummto.dz

Keywords: support vector machines, double hinge loss, rejection, classification cost

Received: March 6, 2014

The objective of this study is to minimize the classification cost using Support Vector Machines (SVMs) Classifier with a double hinge loss. Such binary classifiers have the option to reject observations when the cost of rejection is lower than that of misclassification. To train this classifier, the standard SVM optimization problem was modified by minimizing a double hinge loss function considered as a surrogate convex loss function. The impact of such classifier is illustrated on several discussed results obtained with artificial data and medical data.

Povzetek: Predstavljena je optimizacija cene klasificiranja z metodo strojnega učenja SVM.

1 Introduction

Support Vector Machines (SVMs) are becoming one of the most popular pattern recognition schemes due to their remarkable generalization performance. This is motivated by the application of Structural Risk Minimization principle [1, 2]. Because of their good performance in terms of accuracy and generalization, SVMs are frequently used in very complex two-class classification problems.

Even though the generalization performance of support vector classifiers, misclassifications cannot be completely eliminated and, thus, can produce severe penalties. The expected error of a prediction is a very relevant point in many sensitive applications, such as medical diagnosis or industrial applications.

To improve the reliability of classifiers, new machine learning algorithms have been introduced such us conformal prediction determining levels of confidence [3]. Hence, classifications with less confidence than a given threshold may be rejected. This also motivates the introduction of a reject option in classifiers, by allowing for a third decision Ⓡ (Reject) when the conditional probability that an example belongs to each class is close to $1/2$.

Rejecting ambiguous examples has been investigated since the publications of [4, 5] on the error reject tradeoff. A notable attempts to integrate a reject rule in SVMs has been presented in [6]. The authors developed an SVM whose reject region is determined during the training phase. They derived a novel formulation of the SVM training problem and developed a specific algorithm to solve it. Some works have proposed rejection techniques using two thresholds on the output of the SVM classifier and produce a reject region delimited by two parallel hyperplane in the feature space [7, 8]. Other works used mixture of classifiers [9]. This approach is computationally highly expensive.

Recently, some remarkable works have proposed SVM classifier with a reject option using a double hinge loss function. This option was proposed in [10, 11, 12, 13]. The formulation in [10, 11, 12] is restricted to symmetric losses. In [13], the authors have proposed a cost-sensitive reject rule for SVM using an asymmetric double hinge loss. This formulation is based on probabilistic interpretation of SVM published in [14, 15] providing accurate estimation of posterior probabilities. It also generalizes those suggested in [11, 12] to arbitrary asymmetric misclassification and rejection costs. In all these model classifiers, the reject region is defined during the training phase.

In this paper, we develop the training criterion for a generalized SVM with a double hinge loss and then compare the performance of symmetric and asymmetric classification. The optimal classification cost and the error-reject tradeoff have been highlighted through several illustrated tests.

Note that the minimal classification cost must correspond to a good error-reject tradeoff. It is desirable that most of the rejected patterns would have been erroneously classified by the ideal Bayes classifier.

The remainder of this paper is structured as follows. After problem setting in section 2, section 3 recalls Bayes rule with rejection. In section 4, SVM classifier with rejection is developed using the generalized double hinge loss function. After this, the training criterion is detailed. In Section 5, the implementation is tested empirically. Il shows results comparing the considered classifiers. Finally, Section 6 briefly concludes the paper.

2 Problem setting

Let us consider a binary classification problem in which each example belongs to one of two categories. A discriminant $f : \mathcal{X} \mapsto \mathbb{R}$ classifies an observation $x \in \mathcal{X}$ into one of two classes, labeled +1 or -1 . Viewing $f(x)$ as a proxy value of the conditional probability $P = \mathbb{P}(Y = 1|X = x)$, one is less confident for small values of $|\ f(x)\ |$ corresponding to P around $1/2$. The strategy used in this work is to report $sgn(f(x_i)) = +1$ or -1 if $|f(x_i)|$ exceeds a threshold δ_i and no decision otherwise.

In binary problems, the two types of errors are:

- FP: False Positive, where examples labeled -1 are categorized in the positive class, incurring a loss C_n

- FN: False Negative, where examples labeled $+1$ are categorized in the negative class, incurring a loss C_p.

We also assume that the decision ® incurs a loss, R_n and R_p for rejected examples labeled -1 and $+1$, respectively. This formulation corresponds to [13] . For symmetric classification [10, 11, 12], we have $C_p = C_n = 1$ and $R_p = R_n = r$ with $0 \leq r \leq 1/2$. The expected losses pertaining to each possible decision d are displayed in Figure 1, assuming that all costs are strictly positive. The lower risk $\mathcal{R}$ is:

$$\begin{aligned} \mathcal{R}(d) = min\{C_pP(x), C_n(1 - P(x)), \\ R_pP(x) + R_n(1 - P(x))\} \end{aligned} \tag{1}$$

where $P(x)$ denotes $P(Y = 1|X = x)$. According to (1), one can see in Figure 1 that rejecting a pattern is a viable option if and only if the point G is located above the segment AB. In other terms, if and only if $\frac{R_p}{C_p} + \frac{R_n}{C_n} < 1$ corresponding to $0 \leq r \leq 1/2$ in [10, 11, 12].

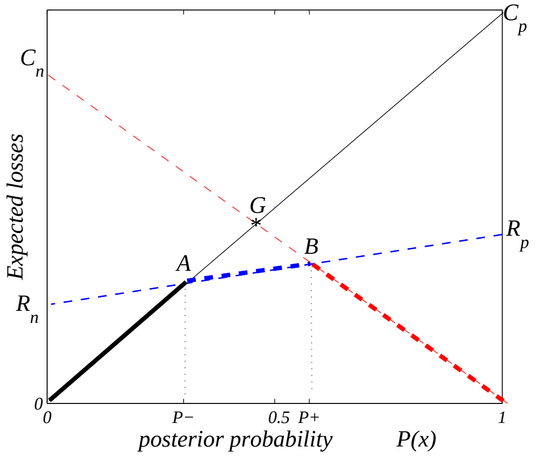


Figure 1: Expected losses against posterior probabilities

3 Bayes rule with rejection

From Figure 1, we deduce that Bayes classifier d^* defined as the minimizer of the risk $\mathcal{R}(d)$ can be expressed simply, using two thresholds:

$$P_+ = \frac{C_n - R_n}{C_n - R_n + R_p} \ , \tag{2}$$

$$P_- = \frac{R_n}{C_p - R_p + R_n} \ , \tag{3}$$

corresponding to symmetric thresholds $P_- = r$ and $P_+ = 1 - r$ in [10, 11, 12].

As Bayes decision rule is defined by conditional probabilities, many classifiers first estimate the conditional probability $\widehat{P}(Y = 1|X = x)$, and then plug this estimate in Eq.4 to build the decision rule.

$$f^*(x) = \begin{cases} +1 & \text{if } \ \widehat{P}(Y = 1|X = x) > P_+ \ , \\ -1 & \text{if } \ \widehat{P}(Y = 1|X = x) < P_- \ , \\ 0 & \text{otherwise} \ . \end{cases} \tag{4}$$

where $f^*(x)$ corresponds to the decision d^*, minimizer of the risk (1).

4 SVM classifier with Reject option (SVMR)

To minimize the empirical counterpart of the risk (1) computationally not feasible, one could replace it by surrogate loss functions. The most popular are the hinge loss motivated by [1] leading to sparse solutions [13, 12] and the logistic regression model offering ability to estimate the posterior probability $\widehat{P}(Y = 1|X = x) = 1/(1 + exp(-yf(x)))$ and then a good choice of the thresholds δ_i. In this study, $\widehat{P}(Y = 1|X = x)$ have to be accurate only in the neighborhood of P_+ and P_- (see equation 4).

4.1 Double hinge loss

The generalized double hinge loss introduced in [13] is a convex and piecewise linear loss function that is tangent to the negative log-likelihood loss at $\delta_+ = log(P_+/(1-P_+))$ and at $\delta_- = log(P_-/(1 - P_-))$ (see Figure 2). This proposal retains the advantages of both loss functions mentioned above: the sparsity of the hinge loss and the ability of the neg-log-likelihood loss to estimate the posterior probability P_+ and P_-, respectively at the tangency points δ_+ and δ_-. So the decision rule can be expressed as:

$$f(x) = \begin{cases} +1 & \text{if } \ f(x) > \delta_+ \ , \\ -1 & \text{if } \ f(x) < \delta_- \ , \\ 0 & \text{otherwise} \ . \end{cases} \tag{5}$$

These thresholds are symmetric in [10, 11, 12], $\delta_+ = -\delta_- = \delta_o$ and the recommended value of δ_o belongs to the interval $[r, 1 - r]$. To express the generalized double

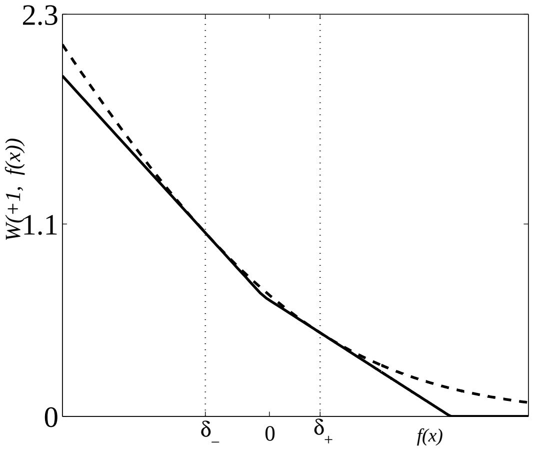


Figure 2: *Double hinge loss function for positive examples, with $P_- = 0.35$ and $P_+ = 0.6$ (solid: double hinge, dashed: likelihood)*

hinge function [13], we consider firstly the standard logistic regression procedure where φ is the negative log-likelihood loss:

$$\varphi(y, f(x)) = log(1 + exp(-yf(x))) \ . \qquad (6)$$

that is $\varphi(+1, f(x)) = log(1 + exp(-f(x)))$ for positive examples and $\varphi(-1, f(x)) = log(1+exp(f(x)))$ for negative examples. Let us work on Figure 3 corresponding to positive examples ($y_i = +1$).

$W = a_1 f(x) + g_1$ is the first slop (right to left) of $W(+1, f(x))$ where $a_1 = \frac{d[\varphi(+1,f(x))]}{d[f(x)]}|_{\delta_+} = -(1 - P_+)$.

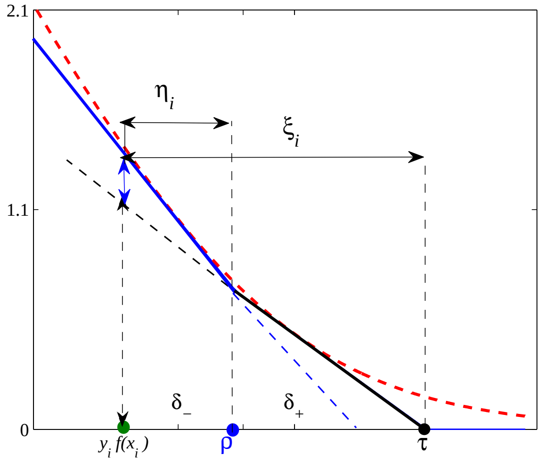


Figure 3: *Double hinge loss function for positive examples, with $P_- = 0.4$ and $P_+ = 0.7$ (solid: double hinge, red dashed: likelihood)*

At the tangency point δ_+, we have $\varphi(+1, f(x)) = W(+1, f(x))$, hence $g_1 = -p_+ log(P_+) - (1-P_+) log(1-P_+) = H(P_+)$.

The second slop of $W(+1, f(x))$ is $W = a_2 f(x) + g_2$ where $a_2 = \frac{d[\varphi(+1,f(x))]}{d[f(x)]}|_{\delta_-} = -(1 - P_-)$ and $g_2 = -P_- log(P_-) - (1 - P_-) log(1 - P_-) = H(P_-)$.

For $a_1 f(x) + g_1 = 0$, we have $f(x) = \tau_+ = \frac{H(P_+)}{1-P_+}$ and for $a_1 f(x) + g_1 = a_2 f(x) + g_2$, we have $f(x) = \rho_+ = \frac{H(P_-)-H(P_+)}{P_+-P_-}$. The double hinge function for positive examples is then expressed as:
$W(+1, f(x)) =$

$$\begin{cases} -(1 - P_-)f(x) + H(P_-) & \text{if } f(x) < \rho_+ \\ -(1 - P_+)f(x) + H(P_+) & \text{if } \rho_+ \leq f(x) < \tau_+ \\ \ \ 0 & \text{otherwise,} \end{cases} \qquad (7)$$

The same strategy of calculation leads to the double hinge function for negative examples: $W(-1, f(x)) =$

$$\begin{cases} P_+ f(x) + H(P_+) & \text{if } f(x) > \rho_- \ , \\ P_- f(x) + H(P_-) & \text{if } \tau_- \geq f(x) > \rho_- \\ \ \ 0 & \text{otherwise.} \end{cases} \qquad (8)$$

where $\tau_- = \frac{-H(P_-)}{P_-}$ and $\rho_- = \rho_+ = \rho$. The double hinge loss ψ_r introduced in [10, 11, 12] is a scaled version of the loss W. It is given by $\psi_r(yf(x)) =$

$$\begin{cases} 1 - \frac{1-r}{r} yf(x) & \text{if } yf(x) < 0 \\ 1 - yf(x) & \text{if } 0 \leq yf(x) < 1 \\ \ \ 0 & \text{otherwise} \end{cases} \qquad (9)$$

hence,

$$\psi_r(yf(x)) = \frac{1}{H(r)} W \left(y, \frac{H(r)}{r} f(x) \right) \ .$$

where $H(r) = H(P_-)$ and $H(P_-) = H(P_+)$ in the symmetric case. Note that, although minimizing $\psi_r(yf(x))$ or W will lead to equivalent solutions for f. With minimizing $\psi_r(yf(x))$, the decision rule recommended by [11] classifies an example when $|f(x)| > \delta_o = \frac{1}{2}$, while in [13], an example is classified when $|f(x)| > \frac{r}{H(r)} log \frac{r}{1-r}$. The last decision rule rejects more examples when the loss incurred by rejection is small and fewer examples otherwise. The two rules are identical for $r = 0.24$.

4.2 Training Criterion

As in standard SVMs, we consider the regularized empirical risk on the training sample. Introducing the double hinge loss (7-8) results in an optimization problem that is similar to the standard SVMs problem.

4.2.1 Primal problem

Let C_o a constant to be tuned by cross-validation, we define $D = C_o(P_+ - P_-)$, $B_i = C_o(1 - P_+)$ for positive examples and $B_i = C_o P_-$ for negative examples. The primal optimization problem reads

$$\begin{aligned} min_{f,b} \ & \tfrac{1}{2}\|f\|^2_{\mathcal{H}} + \\ & \textstyle\sum_{i=1}^n B_i \left|\tau_i - y_i(f(x_i) + b)\right|_+ + \\ & D \textstyle\sum_{i=1}^n \left|\rho - y_i(f(x_i) + b)\right|_+ \end{aligned} \qquad (10)$$

where $|\cdot|_+ = max(\cdot, 0)$. The (squared) norm of f is a regularization functional in a suitable Hilbert space. The primal problem (10) is best seen with introduction of slack variables ξ and η shown in Figure(3).

$$\left\{\begin{array}{lll} min_{f,b,\xi,\eta} & \frac{1}{2}\|f\|^2_{\mathcal{H}} + \sum_{i=1}^{n} B_i\xi_i + D\sum_{i=1}^{n}\eta_i \ , \\ Sc & y_i(f(x_i)+b) \geq \tau_i - \xi_i, \ i = 1,\dots,n \\ & y_i(f(x_i)+b) \geq \rho - \eta_i, \ i = 1,\dots,n \\ & \xi_i \geq 0 \ , \quad \eta_i \geq 0, \qquad i = 1,\dots,n. \end{array}\right. \tag{11}$$

4.2.2 Dual problem

The Lagrangian of (11) is given by:

$$\left\{\begin{array}{l} L(f,b,\xi,\eta,\alpha,\beta,\upsilon,\omega) = \frac{1}{2}\|f\|^2 + \sum_{i=1}^{n} B_i\xi_i \\ +D\sum_{i=1}^{n}\eta_i - \sum_{i=1}^{n}\alpha_i\left[y_i(f(x_i)+b) - \tau_i + \xi_i\right] \\ -\sum_{i=1}^{n}\beta_i\left[y_i(f(x_i)+b) - \rho + \eta_i\right] \\ -\sum_{i=1}^{n}\upsilon_i\xi_i - \sum_{i=1}^{n}\omega_i\eta_i \end{array}\right. \tag{12}$$

with:
$\upsilon_i \geq 0, \ \omega_i \geq 0, \ \alpha_i \geq 0, \ \beta_i \geq 0, \ and\, i = 1,\dots,n.$

The Kuhn-Tucker conditions imply:

$$\left\{\begin{array}{lll} \frac{\partial L}{\partial b} = 0 & \Rightarrow & \sum_{i=1}^{n}(\alpha_i + \beta_i)y_i = 0 \\ \frac{\partial L}{\partial f} = 0 & \Rightarrow & \sum_{i=1}^{n} f(\cdot) = (\alpha_i + \beta_i)y_i k(\cdot, x_i) \\ \frac{\partial L}{\partial \xi_i} = 0 & \Rightarrow & B_i - \upsilon_i - \alpha_i = 0 \Rightarrow 0 \leq \alpha_i \leq B_i \\ \frac{\partial L}{\partial \eta_i} = 0 & \Rightarrow & D - \omega_i - \beta_i = 0 \Rightarrow 0 \leq \beta_i \leq D \end{array}\right. \tag{13}$$

for $i = 1,\dots,n$. Thanks to these conditions, we can eliminate f, ξ and η from the Lagrangian.

$$\left\{\begin{array}{ll} L(\alpha,\beta) = & \frac{1}{2}(\alpha+\beta)^T G(\alpha+\beta) - \tau^T\alpha - \rho^T\beta \\ Sc & y^T(\alpha+\beta) = 0 \\ & 0 \leq \alpha_i \leq B_i, i = 1,\dots,n \\ & 0 \leq \beta_i \leq D, i = 1,\dots,n \end{array}\right. \tag{14}$$

where $\tau = (\tau_1,\dots,\tau_n)^T$ et $\rho = (\rho_1,\dots,\rho_n)^T$ are the threshold vectors of $\mathbb{R}^n$, G is the $n \times n$ influence matrix with general term $G_{ij} = y_i y_j k(x_i, \ x_j)$ and $k(.,.)$, is the reproducing kernel of the Hilbert space $\mathcal{H}$. Let $\gamma = \alpha + \beta$, the problem (14) can be rephrased as:

$$\left\{\begin{array}{lll} max_{\alpha,\gamma} & -\frac{1}{2}\gamma^T G\gamma + (\tau-\rho)^T\alpha + \rho^T\gamma & , \\ Sc & y^T\gamma = 0 & , \\ & 0 \leq \alpha_i \leq B_i, \qquad i = 1,\dots,n & , \\ & 0 \leq \gamma_i - \alpha_i \leq D, \quad i = 1,\dots,n & \end{array}\right. \tag{15}$$

The problem (15) is a quadratic problem under box constraints. Compared to the standard SVM dual problem, one has an additional vector to optimize, but we will show that α is easily recovered from γ.

4.2.3 Solving the problem

To solve the dual (15), the strategy used in the active set method [17] is considered. Firstly, the training set is partitioned in support and non support vectors. the training criterion is optimized considering this partition. Then, this optimization results in an updated partition of examples in support and non-support vectors. These two steps are iterated until predefined level of accuracy is reached. Table (1) shows how the training set is partitioned into five subsets defined by the constraints in (15).

The outcomes of the membership of example i to one of the subsets described above has the following consequences on the dual variables (α, γ):

$$\left\{\begin{array}{lll} i \in I_0 \ \Rightarrow \alpha_i = 0 & \gamma_i = 0 & ; \\ i \in I_\tau \ \Rightarrow 0 \leq \alpha_i \leq B_i & \gamma_i = \alpha_i & ; \\ i \in I_B \Rightarrow \alpha_i = B_i & \gamma_i = B_i & ; \\ i \in I_\rho \ \Rightarrow \alpha_i = B_i & B_i < \gamma_i < B_i + D & ; \\ i \in I_D \Rightarrow \alpha_i = B_i & \gamma_i = B_i + D & . \end{array}\right. \tag{16}$$

Hence, provided that the partitioning is known, γ_i has to be computed only for $i \in I_\tau \cup I_\rho$. Furthermore, α_i is either constant or equal to γ_i.

We saw that, assuming that the examples are correctly partitioned, problem 15 can be solved by considering a considerably smaller problem, namely the problem of computing γ_i for $i \in I_\tau \cup I_\rho$. Let $I_c = \{I_B, I_D\}$ and $I_h = \{I_\tau, I_\rho\}$. The problem (15) becomes:

$$\left\{\begin{array}{l} L(\gamma) = \quad \frac{1}{2}\gamma^T G\gamma - (S)^T\gamma \\ Sc: \quad y^T\gamma = 0 \\ 0 \leq \gamma_i \leq C, \ i = 1,\dots,n \, and \, C_i = B_i + D \end{array}\right. \tag{17}$$

The relation between the parameters of the preceding formulation and the initial parameters of the problem (11) can be obtained after formulating the Lagrangian of the dual (17)

$$\left\{\begin{array}{l} L(\gamma,\lambda,\mu,\nu) = \\ \frac{1}{2}\gamma^T G\gamma - S^T\gamma + \lambda\gamma^T y - \nu^T\gamma + \mu^T(\gamma - C\mathbb{I}^n) \end{array}\right. \tag{18}$$

where the Lagrange multipliers λ, μ, ν must be positive or null and $\mathbb{I}^n$, a vector of 1. This Lagrangian can be compared with the Lagrangian of the primal (11) reformulated as follows:

$$\left\{\begin{array}{l} L = \frac{1}{2}\|f\|^2 - \sum_{i=1}^{n}\gamma_i y_i f(x_i) - \sum_{i=1}^{n}\gamma_i y_i b \\ +\sum_{i=1}^{n}\alpha_i(\tau - \rho) + \sum_{i=1}^{n}\gamma_i\rho \\ +\sum_{i=1}^{n}\xi_i(B_i - \alpha_i - \upsilon_i) \\ +\sum_{i=1}^{n}\eta_i(D - \beta_i - \omega_i) \end{array}\right. \tag{19}$$

by replacing the variable f by γ, the problem (19) becomes:

$$\left\{\begin{array}{l} L(\gamma,b,\xi,\eta) = \\ \frac{1}{2}\gamma^T G\gamma + b\gamma^T y - S^T\gamma \\ +\xi^T(\alpha + \upsilon - B) + \eta^T(\gamma + \omega - D\mathbb{I}^n) \end{array}\right. \tag{20}$$

I_0	saturated part of the loss	$I_0 = \{i \vert y_i(f(x_i) + b) > \tau\}$
I_τ	first hinge of the loss	$I_\tau = \{i \vert y_i(f(x_i) + b) = \tau\}$
I_B	first slop of the loss	$I_B = \{i \vert \rho < y_i(f(x_i) + b) < \tau\}$
I_ρ	second hinge of the loss	$I_\rho = \{i \vert y_i(f(x_i) + b) = \rho\}$
I_D	second sop of the loss	$I_D = \{i \vert y_i(f(x_i) + b) < \rho\}$

Table 1: Partitioning the training set

To reveal the relations between the primal and dual variables, we will check the KKT conditions stipulating the cancellation of the gradient of the Lagrangian (20) according to the primal variable γ in the different subsets.

Table 2 describes the properties of each set regarding the original variables and the Lagrange multipliers.

4.2.4 Algorithm

Let us assume the repartition in each set (I_0, I_h and I_c) to be known. Only the values of γ belonging to I_h remain unknown, they will then be given by the solution of the following optimization problem whose dimension is lower than initial dimension. After slightly abusing notations, we define: $\gamma_h = \gamma(I_h)$, $y_h = y(I_h)$, $G_{hh} = G(I_h, I_h)$, $c_C = \sum_{(i \in I_B)} B_i y_i + \sum_{(i \in I_D)} C_i y_i$ and
$S_h =$
$(\tau(I_\tau)^T \rho(I_\rho)^T)^T - G(I_h, I_D)(B(I_B)^T D\mathbb{I}(I_D)^T)^T$.
The problem (17) becomes:

$$\begin{cases} L(\gamma_h) = & \frac{1}{2}\gamma_h^T G_{hh} \gamma_h - S_h^T \gamma_h \\ Sc & y_h^T \gamma_h + c_C = 0 \end{cases} , \tag{21}$$

The Kuhn-Tucker conditions gives us the system to be solved to find the values of γ that are still unknown.

$$\begin{cases} G_{hh}\gamma_h = S_h - y_h^T \lambda \\ y_h^T \gamma_h = -c_C \end{cases} , \tag{22}$$

After resolving this system, a component of γ violating the primal or dual constraints must be moved to the suitable set. The process is iterated until all box constraints are satisfied.

During the learning process, the time consuming step is the resolution of the linear system (22). For this, we used the incremental strategy outlined in [18] whose complexity is close to $\mathcal{O}(n^2)$ The presented SVMR computational complexity is comparable to that of the standard SVM [18]. The only computational overhead is that the presented SVMR uses 5 categories of examples while SVM uses three.

5 Results and discussions

Data:
To evaluate the performance of the SVMR classifiers, three types of data have been used:

- synthetic data generated with a classical dataset with two gaussianly distributed classes with similar variances but different means chosen to create many ambiguous examples.
- as medical decision making is an application domain for which rejection is of primary importance, data related to medical problems will be considered. ElectroCardioGram (ECG) records from (www.physionet.org /physiobank/database/mitdb) are used. Each tape is accompanied by an annotation file. in which ECG beats have been labeled by expert cardiologists. Since this study is to evaluate the performance of a binary classifier with a reject option, we followed the AAMI recommended practice [19] to form two heartbeat classes: (i) the positive class representing the ventricular ectopic beats (V); (ii) the negative class representing the normal beats (N), including Normal beats, Left Bundle Branch Block beats (LBBB) and Right Bundle Branch Block beats (RBBB). In agreement with [19], records containing paced beats (102, 104, 107, 217) and 23 records with no V beat or less than 40 V beat were excluded leaving 21 records of interest. We have stored each beat by a 7-feature vector. The feature extraction is described in [20]
- For experimenting with large data, the forest CoverType database from UCI repository was also used. (http://kdd.ics.uci.edu/databases/covertype/). We consider the subproblem of discriminating the positive class Cottonwood (2747 examples) against the negative class Douglas-fir (17367 examples).

Tests:
The first series of experiments are done with the ECG data to explain the effectiveness of the classification with rejection. We selected record 214 and 221 containing together 3546 of N beats and 652 of V beats. As no cost matrix is provided with this data, we assume that $R_p = R_n = r$ as in [10, 11, 12] and $P_+ = 1 - \frac{C_p}{C_n} P_- = 1 - \theta P_-$ where $\theta = 1$ in [10, 11, 12] and $\theta \geq 1$ in [13]. Often, in practice, especially in medical applications, FN errors are more costly than FP errors ($\theta > 1$). Figures 4 and 5 show respectively an example of the reject region produced by the SVMR classifier for $\theta = 1$ and for $\theta > 1$. In Figure 5, the SVMR classifier encourages the rejection of more FN examples because they are more costly than FP examples.

All previous classifiers comparatives studies have been based on the error rates obtained, but error rate is not the

Set	Initial constraints	Primal constraints	Dual constraints
I_0	$y_i[f(x_i)+b] > \tau_i$	$\xi_i = \eta_i = 0$	$\mu = 0, \nu = G\gamma + by - \tau \neq 0$
I_τ	$y_i[f(x_i)+b] = \tau_i$	$\xi_i = \eta_i = 0$	$\mu = 0, \nu = G\gamma + by - \tau = 0$
I_B	$\rho < y_i[f(x_i)+b] < \tau_i$	$\xi_i \neq 0, \eta_i = 0$	$\nu = 0, \mu = -G\gamma - by + \tau = \xi$
I_ρ	$y_i[f(x_i)+b] = \rho$	$\xi_i \neq 0, \eta_i = 0$	$\nu = 0, \mu = G\gamma + by - \rho = 0$
I_D	$y_i[f(x_i)+b] < \rho$	$\xi_i \neq 0, \eta_i \neq 0$	$\nu = 0, \mu = -G\gamma - by + \rho = \eta$

Table 2: Situation of the constraints for the five types of examples

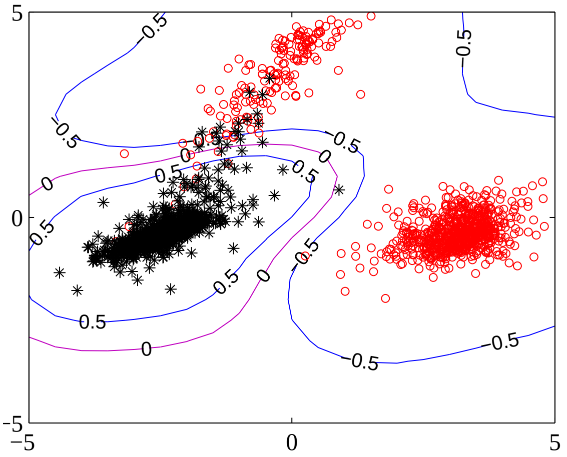


Figure 4: Scatter plot showing the reject region induced by the reject thresholds in correspondence to the costs of misclassifying and rejecting samples. Positive cases are represented by black asts and negative cases by red circles. The lines $+0.5$ and -0.5 correspond respectively to δ_o and $-\delta_o$ and the line 0 corresponds to $f(x) = 0$ or $P(Y = 1 \mid X = x) = 0.5$.

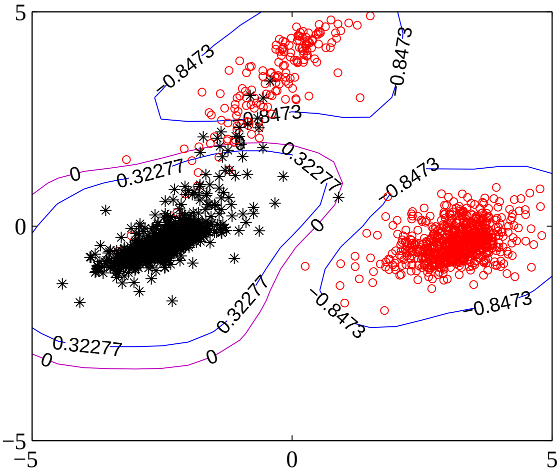


Figure 5: Scatter plot showing the reject region induced by the reject thresholds in correspondence to the costs of misclassifying and rejecting samples. Positive cases are represented by black asts and negative cases by red circles. The lines $+0.32277$ and -0.8473 correspond respectively to δ_+ and δ_- and the line 0 corresponds to $f(x) = 0$ or $P(Y = 1 \mid X = x) = 0.5$.

only measurement that can be used to judge a classifier's performance. In many applications, the classification cost is a parameter witch will be considered since Bayes classifiers with or without rejection aim to minimize the classification cost.

For illustration, we compare the reject rates obtained with the SVMR classifiers proposed in [10, 11, 12] where the reject threshold $\delta_o \in [r, \quad 1-r]$ and the SVMR classifier proposed in [13] where the reject thresholds are $\delta_+ = \log(P_+/(1-P_+))$ and $\delta_- = \log(P_-/(1-P_-))$ respectively for positive and negative examples. For this purpose, we consider the symmetric classification, $P_+ = 1 - P_-$. Figure 6 and 7 obtained with synthetic data and ECG data (record 214 and 221) show that in all cases, the decision rule [13] rejects fewer examples when the loss incurred by rejection is high and more examples otherwise. The rule in [10, 11, 12] considers the reject threshold $\delta_o = 1 - r$ as the largest value of δ_o and then rejects more examples for all reject costs. For $\delta_o = r$, this rule rejects less frequently especially when r close to zero, it becomes almost with no rejection. For the middle value $\delta_o = 0.5$ seen as a compromise among r and $1-r$, the rule [10, 11, 12] and the one proposed in [13] are identical at $r = 0.24$.

As pointed out in [5], the advantage of classifying with rejection can be judged by the error-reject tradeoff. Since the error rate E and the reject rate R are monotonic functions of r. We compute the tradeoff E versus R from $E(r)$ and $R(r)$ when r varies between 0.5 and 0.12 and the threshold $\delta_o = 0.5$ recommended in [11, 12]. Figure 8 shows the error reject tradeoff for the rule proposed in [13] (black curves) and for the rule proposed in [10, 11, 12] (red curves). The obtained results differ due to the size of the rejection region induced by the rules. From these results, we can conclude another interesting parameter that is the error-reject ratio defined in [5] that is $\frac{\triangle E}{\triangle R}$ (dashed lines). For high reject costs $(0.4 \leq r \leq 0.5)$, the rule [13] indicates an error-reject ratio of -0.58, -0.84 and -0.42 respectively for synthetic data, ECG data and forest data. This means that 58%, 84% and 42% respectively of the rejected patterns would have been erroneously classified. Using the rule proposed in [10, 11, 12] with $\delta_o = 0.5$, the error-reject ratios obtained are -0.15 for synthetic data, -0.23 for ECG

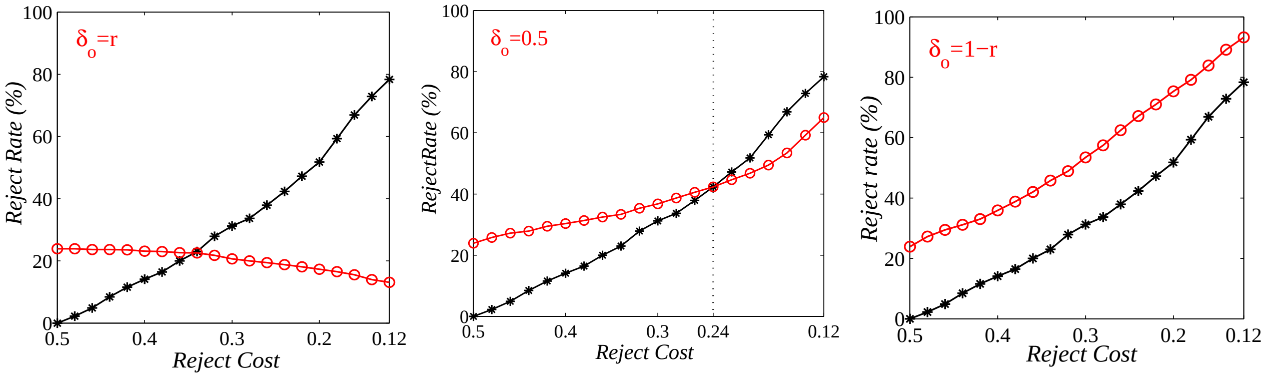


Figure 6: Comparison of the reject rate versus the reject cost r obtained with the SVMR in [13] (black curves) and with the SVMR introduced in [10, 11, 12] (red curves). These results are obtained with synthetic data.

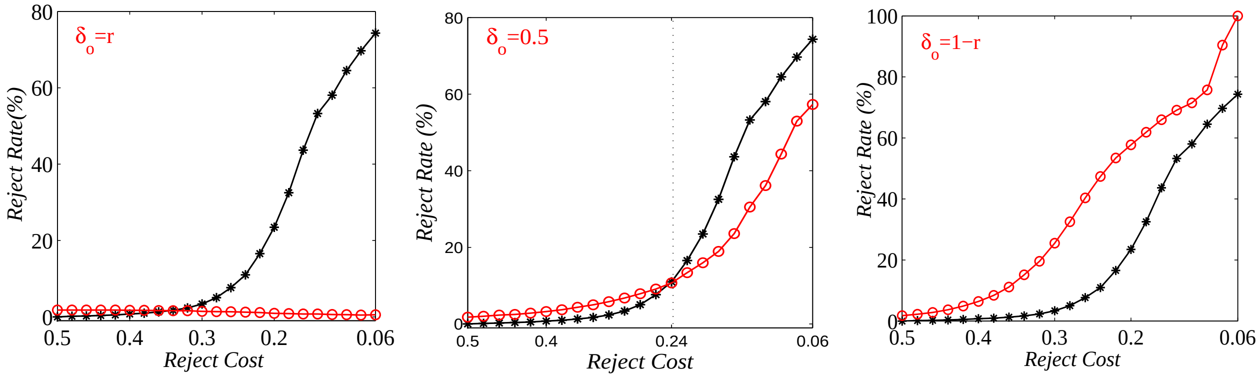


Figure 7: Comparison of the reject rate versus the reject cost r obtained with the SVMR in [13] (black curves) and with the SVMR introduced in [10, 11, 12] (red curves). These results are obtained with medical data.

data and -0.22 for forest data. This means that only 15%, 23% and 22% respectively of the rejected patterns would have been erroneously classified. Hence, it is clear that the rule [13] should lead to a better classification cost.

The last series of tests was carried out using all the selected ECG records. The mean results obtained are reported in Figures 9 and 10. The error against the reject decreases until a quasi constant rate (Figure 9) . Another interesting plot in the same figure represents the error reject ratio. The inflection point in this plot is interesting since it indicates the most important variation of the error against the variation of the reject rate. Two statistical parameters are also used to highlight the performance of the reject rule [13]. The sensitivity and positive predictivity are computed by

$$S_e = \frac{TP}{TP+FN}; \qquad P_p = \frac{TP}{TP+FP}$$

where True Positive (TP) are the samples labeled $+1$ categorized in the positive class. Figure 10 (top) indicates the variation of the classification cost given by

$$C_c = [C_p FN + C_n FP + rR_{rej}]/N_{tot} \qquad (23)$$

where R_{rej} is the number of rejected patterns and N_{tot}, the total number of examples. The same Figure shows that the optimal classification cost C_c corresponds to a good error-reject tradeoff (see Figure 9). Figure 10 (bottom) shows that the positive predictivity is close to 99.8%. In the same figure, it is shown that we obtained more than $98,2\%$ of sensitivity with no rejection and more than 99% of sensitivity for the minimal classification cost with rejection considering $R_p = R_n = r$ and $C_n = 1$ and $\theta = 1.2$. In the same figure, it is clearly shown that the optimal classification cost is not obtained for r=0.5 (simple Bays rule) but for a rejection rate equal to 1.8%. In any application, one must choose the error rate and the rejection rate corresponding to the minimal classification cost. It is the goal of using a cost sensitive classifier.

For a better appreciation of such reject schemes, it should be desirable to perform tests on data accompanied by real cost matrix.

Even though the considered classifier based on sparse probabilistic interpretation of SVM, providing an accurate estimation of posterior probabilities, it should be interesting to assign confidence values to each classification. This can be considered by introducing conformal predic-

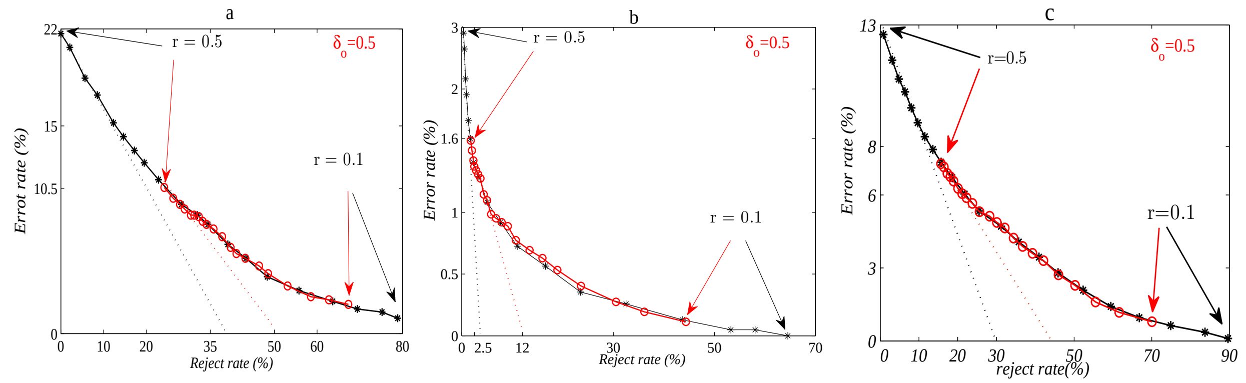


Figure 8: Error versus reject tradeoff obtained using synthetic data (a), ECG data (b) and forest data (c); with [13] (black curves) and with [10, 11, 12] using $\delta_o = 0.5$(red curves).

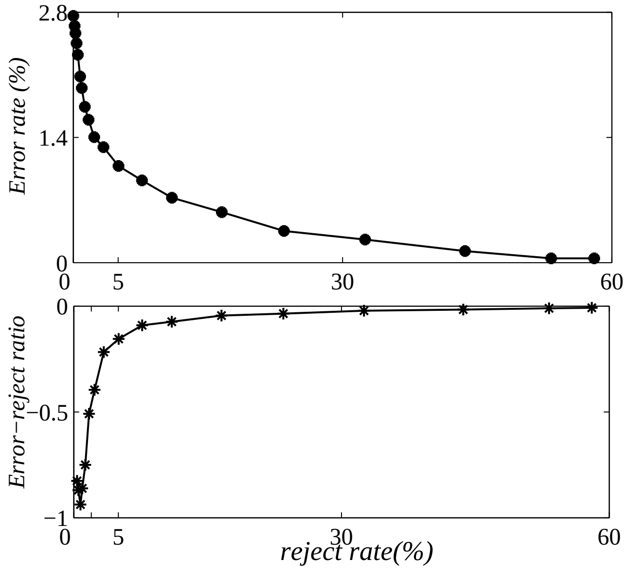


Figure 9: Top: Error rate vs. Reject rate. Bottom: Variation of the error rate against the variation of the reject rate

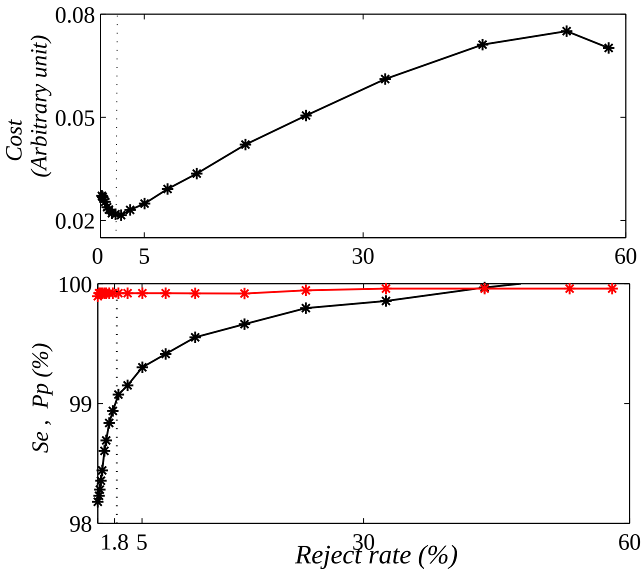


Figure 10: Top: Classification cost against Reject Rate. Bottom: Sensitivity (black curve) and positive predictivity (red curve) against reject rate.

tion whose relationship with rejection is clearly relevant, whether the rejection related to the ambiguity of examples or that related to their atypical characters.

6 Conclusion

This paper presents a cost-sensitive reject rules for SVMs using a double hinge loss. The solution inspired by the probabilistic interpretation of SVM, owns the advantage of the hinge loss function which leads to a consistent solution and the advantage of negative log-likelihood loss which allows a good estimation of posteriori probabilities in the vicinity of the decision thresholds. Note that these dynamic reject thresholds follow the cost of rejecting a sample and the cost of misclassifying a sample. This viewpoint aims to minimize the classification cost.

A possible improvement of this study is to estimate the level of confidence of the classifier by introducing the conformal prediction. This will be a crucial advantage, especially for medical applications, the risk of clinical errors may be controlled by an acceptable level of confidence for a given decision.

References

[1] V. N. Vapnik (1995) The Nature of Statistical Learning Theory *Springer Series in Statistics*

[2] N. Cristianini and J. Shawe-Taylor (2000) An Introduction to Support Vector Machines, *Cambridge University Press.*

[3] Glenn Shafer and Vladimir Vovk (2008) A Tutorial on Conformal Prediction, *Journal of Machine Learning Research*, 9, pp 371-421.

[4] C. K. Chow. (1957) An optimum character recognition system using decision function, *IRE Trans. Electronic Computers*, EC-6(4), pp. 47–254.

[5] C. K. Chow (1970) On optimum recognition error and reject tradeoff, *IEEE Trans. on Information Theory*, 16(41),pp. 41–46.

[6] G. Fumera and F. Roli (2002) Support vector machines with embedded reject option, *In S.-W. Lee and A. Verri, editors, Pattern Recognition with Support Vector Machines: First International Workshop*, volume 2388 of Lecture Notes in Computer Science-Springer, pp. 68–82.

[7] J. T. Kwok (1999) Moderating the outputs of support vector machine classifiers, *IEEE Trans. on Neural Networks*, 10(5), pp 1018–1031.

[8] F. Tortorella. (2004) Reducing the classification cost of support vector classifiers through an ROC-based reject rule, *Pattern Analysis and Applications*, 7(2) pp. 128–143.

[9] A. Bounsiar, E. Grall, P. Beauseroy. (2007) A Kernel Based Rejection Method for Supervised Classification. , *International Journal of Computational Intelligence*, 3(4), pp. 312–321.

[10] R. Herbei and M. H. Wegkamp (2006) Classification with reject option, *The Canadian Journal of Statistics*, 34(4), pp. 709–721.

[11] P. L. Bartlett and M. H. Wegkamp (2008) em Classification with a reject option using a hinge loss, *Journal of Machine Learning Research*, 9, pp. 1823–1840.

[12] M. Wegkamp, M. Yuan (2010) em Classification methods with reject option based on convex risk minimization, *Journal of Machine Learning Research*, (11), pp. 111–130.

[13] Y. Grandvalet, A. Rakotomamonjy, J. Keshet et S. Canu (2009) em Support Vector Machines With a Reject Option, *Advances in Neural Information Processing Systems*, (21), pp. 537–544.

[14] Y. Grandvalet, J. Mariethoz, and S. Bengio (2006) A probabilistic interpretation of SVMs with an application to unbalanced classification. *In Y. Weiss, B. Scholkopf, and J. C. Platt, editors, Advances in Neural Information Processing Systems 18 MIT Press*, pp. 467–474.

[15] P. L. Bartlett and A. Tewari (2007) Sparseness vs estimating conditional probabilities: Some asymptotic results, *Journal of Machine Learning Research*, 8, pp. 775–790.

[16] Z. En-hui and Z. Chao, S. Jian and L. Chen (2011) Cost-sensitive SVM with Error Cost and Class-dependant Reject Cost, *International Journal of Computer Theory and Engineering*, 3(1).

[17] S. V. N. Vishwanathan, A. Smola, and N. Murty (2003) SimpleSVM. *In T. Fawcett and N. Mishra, editors, Proceedings of the Twentieth International Conference on Machine Learning AAAI*, pp. 68–82.

[18] G. Loosli, S. Canu, S. Vishwanathan, and M. Chattopadhay (2005) Boite outils SVM simple et rapide. *Revue d'Intelligence Artificielle RIA*, 19(4), pp.741-767, 2005.

[19] R. Mark and R Wallen, (1987) AAMI-recommended practice: Testing and reporting performance results of ventricular arrhythmia detection olgorithms, *Association for the Advencement of Medical Instrumentation, Tech. Rep. AAMI ECAR*, 1987.

[20] Z. Zidelmal, A. Amirou et A. Belouchrani. (2012) Heartbeat classification using Support Vector Machines (SVMs) with an embedded reject option,*International Journal of Pattern Recognition and Artificial Intelligence (IJPRAI)*, vol,26,no. 1,DOI:10.1142/S0218001412500012.

A Metaheuristic Approach for Propagation-Model Tuning in LTE Networks

Lucas Benedičič and Marko Pesko
Research and Development Department
Telekom Slovenije, d.d., Cigaletova 15, SI-1000 Ljubljana, Slovenia
E-mail: {lucas.benedicic, marko.pesko}@telekom.si

Tomaž Javornik
Communication Systems Department
Jožef Stefan Institute, Jamova cesta 39, SI-1000 Ljubljana, Slovenia

Peter Korošec
Computer Systems Department
Jožef Stefan Institute, Jamova cesta 39, SI-1000 Ljubljana, Slovenia

Keywords: LTE, coverage planning, semi-empirical propagation models, metaheuristics, optimization

Received: November 28, 2013

When deploying a new mobile technology such as LTE, it is crucial to identify the factors that affect the radio network in terms of capacity and quality of service. In this context, network coverage is arguably the single most influential factor. This work presents a metaheuristic-optimization approach that automatically adapts the parameters of a signal-propagation model. The optimization procedure is performed per cell, enabling the calculation of accurate network-coverage predictions. The evaluation of the proposed approach is carried out on two different regions in Slovenia, where Telekom Slovenije, d.d., provides LTE coverage. The results show radio-propagation predictions of improved quality when compared to manual and analytical methods.

Povzetek: Avtorji predstavljajo metahevristično metodo za samodejno optimizacijo parametrov semiempiričnih modelov razširjanja radijskega valovanja. Rezultati številnih poskusov v omrežjih LTE kažejo izboljšano natančnost izračunanih napovedi razširjanja radijskega valovanja.

1 Introduction

One of the primary objectives of radio-coverage planning is to efficiently use the allocated frequency band. To this end, radio-coverage prediction tools are of great importance, as they allow network engineers to test different configurations before physically implementing the changes. However, predicting the radio coverage of a mobile network is a complex task, hence the importance of fast and accurate prediction tools. The precision achieved by the software tool of choice is directly related to the accuracy of the signal-propagation model used. For this reason, signal-propagation models that support configurable parameters are preferred, since they allow the model to adapt to different environments and thus to improve the accuracy of the calculated coverage predictions.

The effectiveness of the decision-making process during radio-network planning is tightly coupled with the precision achieved by the propagation model used. In order to obtain a radio-propagation model that accurately reflects the characteristics of the area covered by the mobile network, the parameters of the signal-propagation model are adjusted using data from field-measurement campaigns. Signal-loss adjustment using this method depends on existing field-measurement data, which are collected in advance for the area covered by the target network cells.

In order to adapt the parameters of a signal-propagation model, mainly analytical approaches were proposed in the related literature [1, 8, 27]. These works confirm the suitability of methods based on least-squares theory for the parameter tuning of signal-propagation models.

In this work, we propose the automatic optimization of the parameters of a signal-propagation model using a metaheuristic approach. The objective of such optimization is four-fold. First, using a stochastic optimization approach, the parameters are optimized in order to reflect the local characteristics of the terrain, thus adapting the model to the local environment of a given network cell. Second, based on a set of field measurements, the automatic optimization of model parameters improves the accuracy of the calculated radio-propagation predictions of each network cell, as well as the radio network as a whole. Third, the proposed metaheuristic method improves, under certain geographical conditions, the results achieved by a traditional linear-least-squares approach [1, 8, 27], the application of which only adapts the linear part of the propagation model, i.e., $y = c + mx$. Fourth, tuning the complete parameter set per network cell shows improved results especially in rugged-terrain areas.

As the working schema for tackling the presented problem, we use PRATO, a parallel framework for coverage planning of cellular networks [5]. The framework flexibility allows for coverage planning and optimization of radio networks in general, and LTE in particular. The optimization component featured by PRATO enables the parallel optimization of several parameters, e.g., the parameters of the signal-propagation model.

The remaining of this paper is organized as follows. Section 2 introduces some principles of radio-propagation prediction and the signal-propagation model used. Section 3 describes the optimization problem involving the parameter-tuning of the signal-propagation model, followed by the performed simulations and the achieved results in Section 4. Finally, in Section 5 we draw some conclusions and give guidelines for further work.

2 Radio-propagation prediction

To calculate the radio-propagation predictions, we use a mathematical model based on the well-known Okumura-Hata formula [14, 22]. Other more accurate methods exist, like the ones based on ray tracing [7, 25]. However, these methods are still inefficient in terms of the computational effort required to achieve satisfying results.

On the other hand, (semi-)empirical methods for radio-propagation predictions give acceptable results within a feasible amount of time. For this reason, they became the industry standard for non-deterministic, signal-propagation calculations [3, 6, 14, 21, 22, 23].

2.1 Signal-propagation model

The Okumura-Hata model has been largely studied and shown to be suitable for predicting the signal propagation of LTE networks [2]. In its primary form, the model distinguishes the distance from the receiver to the transmitter, the frequency used and the effective antenna height, i.e., the antenna height above the receiver's level. These variables are taken into account in order to calculate the path loss in open areas (OA), as described in Equation (1).

$$\begin{aligned} L_{\mathrm{OA}}(x,y,\vec{\beta}) &= \beta_1 + \beta_2 \log(d_{(x,y)}) + \beta_3 \log(H_{\mathrm{A}}) \\ &+ \beta_4 \log(d_{(x,y)}) \log(H_{\mathrm{A}}) - 3.2\,(\log(11.75 \cdot H_{\mathrm{R}}))^2 \\ &+ 44.49 \log(F) - 4.78\,(\log(F))^2\,, \end{aligned} \tag{1}$$

where $\vec{\beta} = (\beta_1, \beta_2, \beta_3, \beta_4)$ are the adaptable parameters of the model, $d_{(x,y)}$ is the distance from the transmitter to the topography point with coordinates (x, y) (expressed in kilometers), H_{A} is the effective antenna height of the transmitter (expressed in meters), H_{R} is the antenna height of the receiver (expressed in meters), and F is the frequency (expressed in MHz).

In this work, as well as in [11], the terrain profile is used for non-line-of-sight (NLOS) determination, i.e., the loss due to an obstacle obstruction in the first Fresnel zone of the transmitter [26]. In such case, additional path-loss factors due to the terrain profile and the Earth shape are added to the original formula, the values of which are calculated as in Equation (2).

$$L_{\mathrm{NLOS}}(x,y) = \sqrt{\left(\alpha K(d_{(x,y)})\right)^2 + E(d_{(x,y)})^2}, \tag{2}$$

where α is the knife-edge diffraction control parameter, the value of which is calculated based on the level of obstruction of the Fresnel zone, $K(d_{(x,y)})$ is the knife-edge diffraction loss (in dB), and $E(d_{(x,y)})$ is the correction due to the Earth sphere, the value of which improves the calculated prediction especially for higher base-station towers and distances over 10 kilometers. All three values depend on the characteristics of the topography point with coordinates (x, y). The euclidean distance between the transmitter and the receiver is intentionally calculated using two-dimensional coordinates due to simplicity and the negligible difference when compared to its three-dimensional counterpart.

In order to adequately predict signal-loss effects due to foliage, buildings and other fabricated structures, a supplementary factor based on the land usage (clutter) is included. This technique is adopted by several propagation models for radio networks, e.g., [1, 3, 19, 22]. Consequently, we introduce an extra term for signal loss due to clutter, thus defining the total model-predicted path loss, which is expressed in dB, as in Equation (3).

$$L(x,y,\vec{\beta}) = L_{\mathrm{OA}}(x,y,\vec{\beta}) + L_{\mathrm{NLOS}}(x,y) + L_{\mathrm{CLUT}}(x,y), \tag{3}$$

where $L_{\mathrm{CLUT}}(x,y)$ represents the clutter loss at the topography point with coordinates (x, y).

2.2 Field measurements

In mobile networks, a moving mobile device (or user equipment, UE) constantly performs cell selection/reselection and handover in order to keep the best possible connection to the network. In this context, the best connection is selected by measuring the signal strength or quality of the neighboring cells. In LTE networks, the UE measures two parameters from the reference signal of the network, namely the Reference Signal Received Power (RSRP) and the Reference Signal Received Quality (RSRQ) [23].

For a certain frequency bandwidth, RSRP measures the average received power over the resource elements that carry cell-specific reference signals. RSRP is applicable in both idle (e.g., waiting for a call) and connected (e.g., during a call) modes. During the procedure of cell selection/reselection in idle mode, RSRP is used. On the other hand, RSRQ is only applicable when the UE is in connected mode.

The radio-coverage calculation involves predicting the network coverage over a certain region, and thus over the UEs within it. Hence, in the first place, we are interested on accurately predicting the best connection the UE would select in idle mode and the RSRP field measurements it uses.

In our case, the field measurements representing the RSRP at a given location were collected using a small truck equipped with the spectrum analyzer Rohde & Schwarz, the functionality of which supports LTE signal analysis. The spectrum analyzer was connected to an external omnidirectional antenna mounted on the roof of the truck, at roughly 2 m above the ground, taking measurements at a rate of 2 Hz. The measurement locations were established using a GPS unit. These GPS-informed locations were tested to be compliant with the 60-meter limit mentioned in [1]. The measurements covered a considerable proportion of the target area, with over 100,000 individual points, collected from more than 30 network cells.

To minimize the deviation in the measured RSRP values, and the impact that small-scale fading has in larger-scale path loss [10], all field measurements were post-processed so that a single value, the median, was calculated for each of the measured locations. The resulting RSRP was then used as the field measurement representing the given location, the resolution of which matches the digital elevation model (DEM) and clutter maps used as input data of the optimization process. Note that the DEM data represents the terrain profile of the geographical area of interest.

3 Parameter tuning of the radio-prediction model

The procedure to adapt the parameters of the mathematical model for each of the cells in the target network involves minimizing the deviation of the radio-propagation prediction compared to a given set of field measurements.

In the context of our work, this means that we have to fine tune the parameters of $L_{\mathrm{OA}}(x, y, \vec{\beta})$, as defined in Equation (3). The adjustable parameters are the elements of vector $\vec{\beta} = (\beta_1, \beta_2, \beta_3, \beta_4) \in \mathbb{R}^4$, namely

β_1 the reference loss or offset;

β_2 the loss slope due to distance of the receiver from the transmitter;

β_3 the loss slope due to height of the transmitter antenna;

β_4 the loss slope due to the combined effect of the distance and height of the antenna.

The parameter tuning is performed per cell to improve local fitting of the radio predictions, being its resulting solution a vector $\vec{\beta}^*$ of the target cell.

The analytical approach for tuning of the radio-prediction model consists of correlating the field measurements with the predicted received-signal values. The new parameter set originates from the minimization of an error criterion. As defined in [15, 27], the minimization criterion is the squared-sum difference between the predicted and the observed RSRP levels, the definition of which is shown later in Equation (6).

As a general rule when applying this approach, only the first two components of the vector $\vec{\beta}$ are adapted, i.e., β_1 and β_2, whereas the values of β_3 and β_4 are kept constant. Therefore, the analytical method consists in fitting only the linear part of the path-loss definition previously presented in Equation (3), i.e.:

$$\Delta L(x, y, \vec{\beta}) = \beta_1 + \beta_2 \log(d_{(x,y)}). \tag{4}$$

The expression in Equation (4) does not take the terrain height into account, which is feasible when the field measurements are taken at a roughly constant height relative to the base station [8, 27]. However, when these heights fluctuate within the coverage area of a cell, the other two parameters, β_3 and β_4, have a considerable effect on the adaptation of the signal-propagation model, as it will be shown in the following sections.

3.1 Differential ant-stigmergy algorithm

In order to adapt the vector $\vec{\beta}$, including its four components, we turn our attention to metaheuristic algorithms in general [24] and swarm intelligence in particular [16]. From this last family of metaheuristics, we have chosen the differential ant-stigmergy algorithm (DASA) [17].

A standalone metaheuristic, the DASA is based on the well-known Ant-Colony Optimization (ACO) [9]. It provides a specialized extension for solving high-dimensional, numerical-optimization problems, whereas the ACO operates on the discrete domain. The DASA represents the search space in a fine-grained discrete form, producing a graph. This graph is then used as the walking paths for the ants, which iteratively improve the temporary best solution.

There are several reasons for choosing the DASA as the optimization algorithm in the context of this problem. First, the benefits of metaheuristic algorithms for solving optimization problems, particularly in the context of radio networks, was demonstrated by numerous authors [4, 12, 15, 18]. Second, in [17], the authors shown the suitability of the algorithm for solving numerical problems, also exhibiting better performance than other swarm-based metaheuristics. Moreover, it has already been successfully applied for tackling an optimization problem in the area of radio networks [4], obtaining competitive results.

The mapping between the parameter-optimization problem and the DASA is as defined in Equation (5).

$$X_a = \{x_1, x_2, x_3, x_4\}, \tag{5}$$

where X_a is the solution vector of ant a during the minimization process, and x_j represents the j-th component of vector $\vec{\beta}$ for the signal-propagation model of a given cell. At the end of every iteration, and after all the ants have created solutions, they are evaluated to establish if any of them is better than the best solution found so far.

For a more in-depth explanation about this procedure and the DASA itself, we refer the reader to [17].

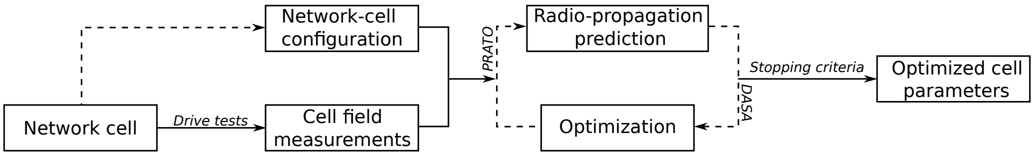


Figure 1: The parameter-optimization system as executed per network cell by a parallel process of PRATO.

3.2 Optimization objective

The optimization objective consists in adjusting the values of components of vector $\vec{\beta}$ according to a set of field measurements of a given cell. Each network cell is independently optimized, so that its radio-propagation prediction minimizes the mean-squared error against the field measurements, as defined in Equation (6).

$$f^*(c) = \min \sum_{m \in M_c} \frac{(p_c - L_c(\mathrm{coord}(m), \vec{\beta}) - m)^2}{|M_c|} \quad \forall c \in N, \qquad (6)$$

where $f^*(c)$ is the optimization objective to be minimized for cell c, N is one of the test networks, p_c is the transmit power of cell c, m is a field measurement of cell c, M_c is the set of all field measurements of cell c, and $L_c(\mathrm{coord}(m), \vec{\beta})$ represents the path loss of cell c at the same geographical point of the field measurement m, as defined in Equation (3). Note that $f^*(c)$ is independently calculated for each $c \in N$.

4 Simulations

In the parameter-optimization problem, the process starts with a mobile network. Each network cell is optimized by an independent parallel process as shown in Figure 1. The set of field measurements corresponding to the cell under optimization has to be gathered through drive tests before hand. Together with the cell configuration, they provide the input data for the optimization process itself. An iteration begins when the DASA generates a solution vector for each of the ants in the colony. The following step involves the evaluation of the solution vector carried by an ant, i.e., one radio-propagation prediction per iteration of the optimization process. The objective-function value is calculated as defined in Equation (6), and sent back to the DASA for it to generate the next set of solutions. The optimization process involves multiple iterations, which are repeated until some stopping criteria are met. Then, the best solution found represents the optimized values of the tuning parameters for the radio-propagation model of the target network cell.

The optimization process is performed by PRATO in parallel over the worker processes, each of which runs independently of the others while optimizing the parameters of one network cell.

Compared with the analytical approach, the solution of which requires solving a linear system of equations, a large number of evaluations is needed for the metaheuristic optimization to converge to a solution. Therefore, it is essential to exploit the parallel nature of PRATO in order to concurrently optimize the parameter sets of multiple cells within the network. Otherwise, such approach would not be feasible, since the time required to reach a reasonable solution would be excessive.

	Number of cells	Calculation radius [km]	Area [km^2]	Field-measurement proportion [%]
Net$_1$	9	16.00	82.90	4.40
Net$_2$	25	16.00	133.47	6.74

Table 1: Some characteristics of the test networks used for the experimental simulations.

4.1 Test networks

The test networks, Net$_1$ and Net$_2$, are subsets of a real LTE network deployed in Slovenia by Telekom Slovenije, d.d. For the path-loss predictions, we were provided DEM and clutter maps of 25 m^2 resolution. A calculation radius around each network cell limited the path-loss prediction to a distance where it is feasible for an UE to connect to a cell, i.e., when the RSRP is greater or equal to -124 dBm [20]. At the same time, this calculation radius provides enough overlap among neighboring cells to calculate the network coverage over the whole region, for which the receiver height was set to 2 m above ground level. Table 1 provides more information about the test networks used, such as the number of network cells, the area surface, and the covering proportion of the collected field measurements in terms of the total area of each test network.

Net$_1$ represents a network deployed over a dominant agricultural area with almost flat terrain, some forests and waters streams. The other network, Net$_2$, is deployed over a hilly terrain mostly covered by forests.

As the stopping criteria for the optimization runs, we fixed the maximum number of iterations to 250, since the algorithm showed an acceptable convergence profile in all runs. Overall, the framework completed 20,000 objective-function evaluations, i.e., 180,000 radio-coverage predictions for Net$_1$ and 500,000 for Net$_2$.

The simulations were carried out on several computing nodes of the DEGIMA cluster [13] at the Nagasaki Advanced Computing Center (NACC) of the Nagasaki University in Japan. The reason for using a high-end computer cluster as DEGIMA is to exploit the parallel nature of PRATO. To this end, groups of 5 and 13 computing

	Manual		Analytical		DASA	
Test network	Mean	Standard deviation	Mean	Standard deviation	Mean	Standard deviation
Net$_1$	5.88496	13.69240	0.00001	12.64708	0.01974	12.13723
Net$_2$	6.52300	14.35561	0.00007	12.30833	0.01372	10.59590

Table 2: Mean and standard-deviation values of the radio-propagation prediction against the field measurements. The values, expressed in dB, are given when using the manual, analytical and DASA approaches in each test network.

nodes were used for executing the simulations of the different problem instances, i.e., Net$_1$ and Net$_2$, respectively.

The computing nodes were connected by a LAN, over a Gigabit Ethernet interconnect. The nodes were equipped with a Linux 64-bit operating system (Fedora distribution). OpenMPI was used as the message passing implementation, version 1.6.1, the binaries of which were manually compiled with the distribution-supplied *gcc* compiler, version 4.4.4.

After some trial-optimization runs, the six parameters that control the way the DASA explores the search space were set to the following values:

– $m = 80$, the number of ants;

– $b = 10$, the discrete base;

– $\rho = 0.2$, the pheromone dispersion factor;

– $s_+ = 0.01$, the global scale-increasing factor;

– $s_- = 0.01$, the global scale-decreasing factor; and

– $\epsilon = 10^{-5}$, the maximum parameter precision.

The trial runs consisted in doubling m from 5 to 640, and verifying the convergence profile and best solution found. The values of the other parameters were left unchanged.

4.2 Result analysis

The mean and standard-deviation values after correlating the field measurements with the radio-propagation prediction of each test network and parameter set are shown in Table 2.

The parameter set used for the "Manual" column of Table 2 was provided by the radio experts of the Radio Network Department at Telekom Slovenije, d.d. These values were calculated based on manual observations and were applied to all the cells in the network. The column labeled as "Analytical" represents the parameter set calculated by the least-squares approach as presented in the related literature [1, 8, 27]. The last column, which is labeled as "DASA", represents the average parameter set calculated by the DASA after 30 independent runs.

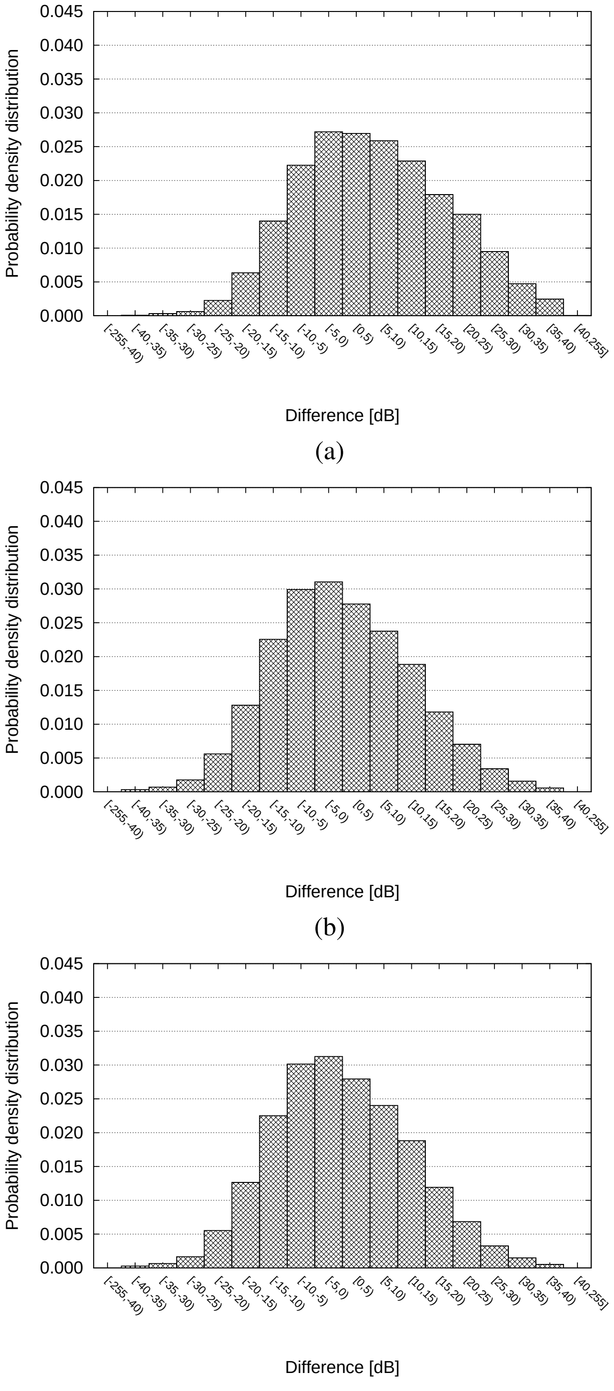
The manual approach clearly shows the biggest discrepancy of the radio-propagation predictions in both the rural (Net$_1$) and hilly (Net$_2$) environments. The analytical approach considerably improves the results achieved by the manual method, thus showing lower mean and standard-deviation values in both test networks. As for the DASA, it further improved the standard deviation of the analytical approach. Moreover, this correction is significant for in Net$_2$, thus confirming the influence of the hilly terrain in the accuracy of signal-propagation predictions. This is especially important on the border of the cell coverage, where a 2 dB difference in the received-signal strength could mean predicting sufficient network coverage where there would otherwise be none. Regarding the mean values showed by the DASA solutions, we may observe that they are several orders of magnitude higher than those of the analytical solutions. However, the values are, in all cases, strictly lower than 0.02 dB, which is a negligible difference in terms of the RSRP levels that outline the coverage of a network cell.

These results confirm that the use of the DASA to perform the optimization of parameters of a signal-propagation model is viable, since it is capable of reflecting the physical phenomena appearing in real-world conditions in two geographically-different network instances.

Figures 2 and 3 depict the probability-density distributions of the difference between the signal-propagation predictions and the field measurements. The mean and standard-deviation values of these distributions are listed in Table 2.

Figure 2 (a) depicts the difference distribution of the coverage prediction for test network Net$_1$ using the manually-calculated parameters, Figure 2 (b) shows the difference distribution for the same test network, but using the analytically-calculated parameters, and Figure 2 (c) shows the difference distribution using the DASA-optimized parameters. Notice how the difference distributions show an improvement when the analytically-calculated parameters are used, lowering the largest (outer) deviations, and raising the lowest (inner) ones. Additionally, the difference is negligible when using the optimized parameters, thus confirming that it is sufficient to only adapt β_0 and β_1 in environments where the height difference between the transmitter and the receiver, i.e., $\log(H_\mathrm{A})$ in Equation (1)), is roughly constant.

The difference distributions of the radio-propagation predictions for test network Net$_2$ using the manually-calculated parameters, the analytically-calculated, and the optimized ones are shown in Figures 3 (a), 3 (b), and 3 (c), respectively. Similar to Net$_1$, the improvement appears in the largest deviations, since their values are lower than when using the manually-calculated parameters. In this case, we may also observe the improvement achieved by the parameter set calculated with the DASA, which included all four components of the vector $\vec{\beta}$, thus better re-

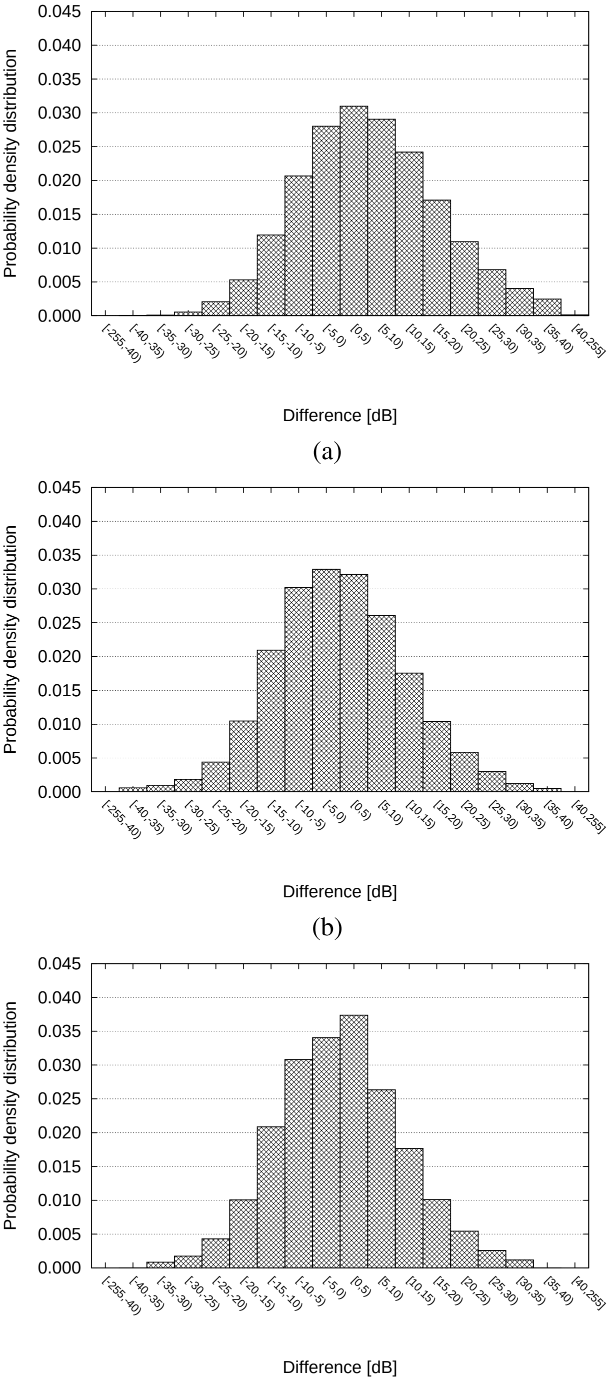


(a)

(b)

(c)

Figure 2: Probability-density distribution of radio predictions against the field measurements of network Net$_1$ over a rural area using the: (a) manually-calculated parameters, (b) analytically-calculated parameters, and (c) optimized parameters.



(a)

(b)

(c)

Figure 3: Probability-density distribution of radio predictions against the field measurements of network Net$_2$ over a hilly area using the: (a) manually-calculated parameters, (b) analytically-calculated parameters, and (c) optimized parameters.

flecting the signal propagation over a hilly terrain.

Overall, the parameter optimization of the signal-propagation model with respect to field measurements does improve the quality of the calculated radio-propagation predictions. Considering that the default parameter values were manually calculated by the radio engineers for the whole network, the convenience of the automated optimization procedure is clear. Indeed, these advantages are a consequence of a simpler method that automatically delivers radio-predictions of superior quality and accurately represent the physical properties of a given environment.

4.3 Statistical analysis

Because of the stochastic nature of the DASA optimization algorithm, we have collected the results of 30 independent runs in order to have enough data for them to be statistically relevant. In other words, the robustness of the results presented in the previous section is analyzed here.

To this end, Table 3 shows a statistical analysis of the mean and standard-deviation values of the solutions reached by the DASA for each test network. The analysis includes the minimum, maximum and average values for every quality measure of the radio-propagation predictions, along with their standard deviation.

We may observe that the standard deviation is consistently lower than 0.015 for both quality measures, indicating a consistent convergence of the optimization algorithm and confirming the suitability of the DASA for tackling the parameter-optimization problem.

5 Conclusion

We have presented a metaheuristic-optimization approach for the parameter optimization of signal-propagation models. The open-source framework for coverage-planning and optimization of radio networks (PRATO)[1] was used to concurrently optimize multiple cells of the target network in parallel, exploiting the resources of a computer cluster. Based on extensive experimental simulations, we have shown the suitability of the metaheuristic approach for the automatic adaptation of the parameter values over different regions of a newly deployed LTE network in Slovenia.

By using different sets of field measurements over the target regions, the combination of the afore-mentioned techniques the parameters of the signal-propagation model were adapted to geographical are around each network cell. As a result, the accuracy of the radio-propagation predictions of the whole network was improved. Moreover, the improvement was significant when applying the presented approach to a network deployed over a hilly area, even outperforming the results of a least-squares analytical method commonly used in the related literature. The simulation results suggest that the presented methodology is applicable for LTE networks in general, since it reached very good accuracy of the calculated radio predictions over diverse terrains. Consequently, the applicability of this approach for arbitrary terrain types can be expected.

Further research will include the performance analysis of other metaheuristic approaches [15] and their result comparison with the DASA. Also, the consideration of urban environments, where the signal-propagation conditions differ from those in rural and hilly areas, should also be explored. Furthermore, in the context of the radio-coverage planning activities carried out at the Radio Network Department of Telekom Slovenije, d.d., supplementary testing of the presented approach, as a support methodology for coverage planning, is currently being conducted. So far, the performed analyses yield robust results compared to traditional, manual techniques.

Acknowledgment

The authors would like to especially thank the staff at the Nagasaki Advanced Computing Center of the Nagasaki University for their support while making the computer cluster DEGIMA available for the experimental simulations. Additionally, the authors are especially thankful to the radio engineers at Telekom Slovenije, d.d., for supplying the network data sets and sharing their professional expertise throughout the creation of this work. This project was co-financed by the European Union, through the European Social Fund.

References

[1] Erik AarnæS and Stian Holm. Tuning of empirical radio propagation models effect of location accuracy. *Wireless Personal Communications*, 30(2-4):267–281, 2004.

[2] Yassir A. Ahmad, Walid A. Hassan, and Tharek A. Rahman. Studying different propagation models for LTE-A system. In *International Conference on Computer and Communication Engineering (ICCCE)*, pages 848–853. IEEE, 2012.

[3] Pamela Begovic, Narcis Behlilovic, and Elma Avdic. Applicability evaluation of Okumura, Ericsson 9999 and Winner propagation models for coverage planning in 3.5 GHZ WiMAX systems. In *19th International Conference on Systems, Signals and Image Processing (IWSSIP)*, pages 256–260. IEEE, 2012.

[4] Lucas Benedičič, Mitja Štular, and Peter Korošec. Balancing downlink and uplink soft-handover areas in UMTS networks. In *IEEE Congress on Evolutionary Computation (CEC)*, pages 1–8. IEEE, 2012.

[5] Lucas Benedičič, Felipe A. Cruz, Tsuyoshi Hamada, and Peter Korošec. A GRASS GIS parallel module

[1] The source code is available for download from the corresponding author's home page, http://cs.ijs.si/benedicic.

	Net$_1$				Net$_2$			
Measure	Min	Max	Avg	Std.dev.	Min	Max	Avg	Std.dev.
Mean	0.00276	0.03639	0.01974	0.01300	0.00156	0.03046	0.01372	0.00846
Standard deviation	12.13657	12.13843	12.13723	0.00057	10.59571	10.59628	10.59590	0.00015

Table 3: Statistical-analysis values of the optimization solutions for each test network. All values are expressed in dB.

for radio-propagation predictions. *International Journal of Geographical Information Science*, 28(4):799–823, 2014.

[6] Dieter J. Cichon and Thomas Kürner. *Propagation Prediction Models*, chapter 4, pages 115–208. European Community, 1995. Available from: http://www.lx.it.pt/cost231/final_report.htm.

[7] Yoann Corre and Yves Lostanlen. Three-dimensional urban EM wave propagation model for radio network planning and optimization over large areas. *IEEE Transactions on Vehicular Technology*, 58(7):3112–3123, 2009.

[8] Chhaya Dalela, Marehalli Prasad, and Pankaj Dalela. Tuning of COST-231 Hata model for radio wave propagation predictions. *Computer Science and Information Technology (CS & IT), DOI*, 10:255–267, 2012.

[9] Marco Dorigo, Mauro Birattari, and Thomas Stutzle. Ant colony optimization. *IEEE Computational Intelligence Magazine*, 1(4):28–39, 2006.

[10] Sven Dortmund, Alfred Schmidt, and Ilona Rolfes. Measurement based channel model for large concert halls. In *IEEE International Symposium on Antennas and Propagation Society (APSURSI)*, pages 1–4. IEEE, 2010.

[11] Sonja Filiposka and Dimitar Trajanov. Terrain-aware three-dimensional radio-propagation model extension for NS-2. *Simulation*, 87(1-2):7–23, 2011.

[12] Mario García-Lozano, Maria A. Lema, Silvia Ruiz, and Flaminio Minerva. Metaheuristic procedure to optimize transmission delays in DVB-T single frequency networks. *IEEE Transactions on Broadcasting*, 57(4):876–887, 2011.

[13] Tsuyoshi Hamada and Keigo Nitadori. 190 TFlops astrophysical N-body simulation on a cluster of GPUs. In *Proceedings of the 2010 ACM/IEEE International Conference for High Performance Computing, Networking, Storage and Analysis*, pages 1–9. IEEE Computer Society, 2010.

[14] Masaharu Hata. Empirical formula for propagation loss in land mobile radio services. *IEEE Transactions on Vehicular Technology*, 29(3):317–325, 1980.

[15] Lianfen Huang, Xiaoxin Chen, Zhibin Gao, and Hongxiang Cai. Online propagation model correction based on PSO algorithm in LTE SON system. In *International Conference on Anti-Counterfeiting, Security and Identification (ASID)*, pages 1–4. IEEE, 2012.

[16] James Kennedy. *Swarm intelligence*, pages 187–219. Springer, 2006.

[17] Peter Korošec, Jurij Šilc, and Bogdan Filipič. The differential ant-stigmergy algorithm. *Information Sciences*, 192:82–97, 2012.

[18] Samip Malla, Birendra Ghimire, Mark C. Reed, and Harald Haas. Energy efficient resource allocation in OFDMA networks using ant-colony optimization. In *International Symposium on Communications and Information Technologies (ISCIT)*, pages 889–894. IEEE, 2012.

[19] Aleksandar Neskovic and Natasa Neskovic. Microcell electric field strength prediction model based upon artificial neural networks. *AEU-International Journal of Electronics and Communications*, 64(8):733–738, 2010.

[20] Michaela Neuland, Thomas Kürner, and Mehdi Amirijoo. Influence of Different Factors on X-Map Estimation in LTE. In *73rd Vehicular Technology Conference (VTC Spring)*, pages 1–5. IEEE, 2011.

[21] Noman Shabbir, Muhammad T. Sadiq, Hasnain Kashif, and Rizwan Ullah. Comparison of radio propagation models for Long Term Evolution (LTE) network. *arXiv preprint arXiv:1110.1519*, 2011.

[22] I. Simic, I. Stanic, and B. Zirnic. Minimax LS algorithm for automatic propagation model tuning. In *Proceeding of the 9th Telecommunications Forum (TELFOR 2001)*, 2001.

[23] Lingyang Song and Jia Shen. *Evolved cellular network planning and optimization for UMTS and LTE*. CRC Press, 2010.

[24] El-Ghazali Talbi. *Metaheuristics: from design to implementation*. Wiley, 2009.

[25] Andrej Vilhar, Andrej Hrovat, Igor Ozimek, and Tomaž Javornik. Efficient open-source ray-tracing methods for rural environment. In *Proceedings of the 16th WSEAS International Conference on Computers*, pages 15–17, 2012.

[26] Howard Xia, Henry L. Bertoni, Leandro R. Maciel, Andrew Lindsay-Stewart, and Robert Rowe. Radio propagation characteristics for line-of-sight microcellular and personal communications. *IEEE Transactions on Antennas and Propagation*, 41(10):1439–1447, 1993.

[27] Mingjing Yang and Wenxiao Shi. A linear least square method of propagation model tuning for 3G radio network planning. In *Fourth International Conference on Natural Computation*, volume 5, pages 150–154. IEEE, 2008.

Multi-class Image Classification Based on Fast Stochastic Gradient Boosting

Lin Li[1,2], Yue Wu[1] and Mao Ye[1]
[1]School. of Computer Science and Engineering, University of Electronic Science and Technology of China
No.2006, Xiyuan Ave, West Hi-Tech Zone, Chengdu, China
E-mail: lilin200909@gmail.com
[2]Sichuan TOP IT Vocational Institute No.2000, Xixin Ave, West Hi-Tech Zone, Chengdu, China

Keywords: multi-class image classification, fast random gradient boosting, multi-class classifier, machine learning, image data set

Received: December 3, 2013

Nowadays, image classification is one of the hottest and most difficult research domains. It involves two aspects of problem. One is image feature representation and coding, the other is the usage of classifier. For better accuracy and running efficiency of high dimension characteristics circumstance in image classification, this paper proposes a novel framework for multi-class image classification based on fast stochastic gradient boosting. We produce the image feature representation by extracting PHOW descriptor of image, then map the descriptor though additive kernels, finally classify image though fast stochastic gradient boosting. In order to further boost the running efficiency, We propose method of local parallelism and an error control mechanism for simplifying the iterating process. Experiments are tested on two data sets: Optdigits, 15-Scenes. The experiments compare decision tree, random forest, extremely random trees, stochastic gradient boosting and its fast versions. The experiment justifies that (1) stochastic gradient boosting and its extensions are apparent superior to other algorithms on overall accuracy; (2) our fast stochastic gradient boosting algorithm greatly saves time while keeping high overall accuracy.

Povzetek: Predstavljena je primerjava algoritmov za večrazredno klasifikacijo slik.

1 Introduction

With the extensive application of the Internet, search engines have become an important tool for people to obtain information, including image information which is one of the most important and interesting information. Traditional search engines on the Internet, including Google, Bing and Baidu have launched a corresponding image search function, but this kind of searching is mainly operated by the file names or related text information of the images. However, it has obvious limitations such as: file name or related information is not accurately related with the image content. So information retrieval based on image content becomes one of the hottest studies of the image retrieval. image classification is based on the image content-based information retrieval, which is based on visual information. Image classification mainly involves two aspects: One is the image feature representation and coding, on the other hand is a classifier selection.

Haralick etc. [1] first proposed a method for feature representation based on image texture features, which is considering the texture characteristics of the image feature space relations, texture and spectral information and its statistical characteristics. Later, considering rotation, affine and other factors, people gradually propose feature representation methods such as LBP [2], SIFT [3], HOG [4] and etc. Statistical represented feature coding method has been widely used, for example a typical representative of the texture histogram representation (histogram of textons) [5] and bag of words or bag of features [6] coding. In recent years, people also proposed a histogram-based pyramid encoding as PHOG (Pyramid Histogram Of Gradient) [7] and PHOW (Pyramid Histogram Of visual Word) [8]. In order to further improving the discriminative capability of feature descriptors, people propose kernel transformation such as Vedaldi's additive kernel transformation [9] can effectively enhance classification performance.

Several classifiers have been successfully used for image classification such as support vector machines, random forests and so on [10]. Haralick etc. [11] first propose method based on image characteristics, using the linear discriminant maximum and minimum decision rules to classify discrimination on the data set and aerial imagery sandstone micrographs. They obtain more than 80% accuracy rate. Ridgeway etc. [12] introduce method based on the corners with features and weighted K nearest neighbor classifier for image classification. They obtain of 93.6% accuracy rate in the 2716 image data sets, and promote performance of the method to three categories (land, forests and mountains, sunset and sunrise). Chapelle etc. [13] use histogram features and support vector machine classifier to achieve the classification performance of 89% accuracy rate on the Corel14 data set. Foody etc. [14] apply support vector machines for remote sensing image classification. They obtain 93.5% accuracy rate, better than the tree algorithm of 90.3% and discriminant analysis method of 90%

accuracy rate.

This paper presents a framework to enhance the natural image classification performance based on PHOW features representation and fast stochastic gradient, and we obtain more than 99% and 84% accuracy rate on the data set Optdigits and 15-Scenes respectively.

2 Fast stochastic gradient boosting algorithm

2.1 Analysis and comparison of algorithms based on decision tree

Traditional classification and regression trees (Classification and Regression Tree, CART) proposed by Breiman [15] is a simple and effective method, but there are many flaws [16]: 1) because decision tree is based on local optimum principle, this will led to the whole tree is not often global optimal. 2) inaccuracies and abnormal training samples have a great impact on the CART. 3) The imbalances of training sample types also affect CART performance.

Improving and enhancing the performance of classification and regression trees is a valuable question. In recent years, bagging and boosting method is the most effective ways. Bagging method [17] is an autonomous improving method, which is a random subtree building based on subsampling over all training samples to obtain samples.

Bagging method proposed by the Breiman [18] is also based on random forest, which use decision trees as a meta-classifier with independent clustering method (Bootstrap aggregation, Bagging), thus produces different training set to generate each component classifier, and finally determine the final classification results by a simple majority vote.

Extreme random tree [19] is similar to the random forest. The tree pieces are combined to form a multi-classifier, the difference with the random forest mainly involving two sides:

1) Sampling the original training samples with replacement strategy, aiming at reducing bias;

2) Splitting test threshold of each decision tree node is selected at random. Assuming split test expression is $split(x) > \theta$, where x is to be classified samples, $split$ is the test function in the random forest classifier, θ is usually based on a sample of a feature set, and in the extreme random forest classifier, θ is randomly selected.

Boosting method [20] is the method, which is starting from the basic classification tree, though iterative process, wrong classification of data give higher weights to build a new round of classification trees greater emphasising on these error detection data. Final classifier classification is based on the principle of majority voting. Despite boosting method is not accurate in some particular cases. But in most cases, it significantly enhances the classification accuracy [21].

Gradient Boosting proposed by Friedman [22] is further improvement over boosting. Their difference with the traditional approach is to improve every computing in order to reduce losses. In order to eliminate losses, it create a new model in the direction of the gradient to reduce losses so that the gradient can be descent. The big difference with conventional methods is that the new model is created from residual losses of the gradient direction of the model in order to reduce losses. Inspiring by bagging random thoughts of Breiman, Friedman introduced stochastic gradient boosting based on random sub-sampling to obtain training samples [23].

In short, bagging and boosting methods both can be called to vote or integrated approach to generate a set of sub-tree or forests, while classification is according to the sub-tree or forest in the whole set or voting on every tree. The difference is that they generate different sub-tree or forests by different ways.

2.2 Fast stochastic gradient boosting algorithm

Fast stochastic gradient boosting algorithm is shown in the Algorithm 1. where $\pi(i)_1^N$ is the random combinations of set of integers $1, 2, ..., N$, assuming sample size of random down-sampling is $\hat{N} < N$, The corresponding sample result is $(y_{\pi(i)}, x_{\pi(i)})_1^{\hat{N}}$. $F_m(x)$ is for the first m points. L is the loss function, M is the number of weak classifiers, C is class sample, R is the leaf node region, J is a terminal leaf number of nodes, ρ is the optimal weak classifier coefficient, S is the number of samples to detect the error, err and err_{min} are the exit variable, **parallel** refers parallel processing.

Algorithm inputs are training samples, outputs $\tilde{F(x)}$ are the output set of weak classifiers.

The step 1 to 17 of the algorithm is weak classifier training processes consisting of three components: The step 2 is randomized sampling. The steps 3 to 10 is weak classifier training stages. The step 11 to 16 is error detection.

The step 2 obtains training samples by randomly sampling for each weak classifier. The step 3 to 10 is weak classifiers training process by classes in turn, which contains: 1) calculating the loss, the loss for classification problems using deviance loss; 2) by calculating the value of the loss of function in the negative gradient of the current model, which was estimated as a residual;3) training a decision tree classifier based on the basic decision tree;4) updating residuals;5) calculating the optimal weak classifier coefficients; 6) generating a new weak classifiers.

The step 11 to 16 is error detection. Classification training and each error detection are simultaneously. Weak classifiers stop training when the error is less than a certain threshold.

Finally, the step 18 is the results of linear combination of weak classifiers constituting the set of training stochastic gradient boosting tree model.

Algorithm 1 Fast stochastic gradient boosting algorithm.

```
Input:
    training data set : $T = (x_1, y_1), (x_2, y_2), ..., (x_N, y_N), x_i \in R^N, y \in Y \in R, N$ is the number of training samples.
    initialization : $F_0(x) = 0, M = 100, err = 0, err_{min} = 0.0001$.
Output:
    combination set of classification trees : $\hat{F}(x)$
    .
for $m = 1$ to $M$ do
    random sampling of (parallel): $\pi(m)^{\hat{(N)}_1} = rand\_perm(m)_1^N$
    for $k = 0$ to $C$ do
        calculating loss : $p_k(x) = \exp(F_k(x))/\sum_{l=1}^{C}\exp(F_l(x)), k = 1, 2, ..., C$
        calculating the gradient (parallel):
            $$\tilde{y}_{\pi(i)k} = -\left[\frac{\partial L(\{y_{\pi(i)l}, F_l(x_{\pi(i)})\}_{l=1}^{C})}{\partial F_k x_{\pi(i)}}\right]_{\{F_l(x)=F_{l,m-1}(x)\}_1^Y} = y_{\pi(i)k} - p_{k,m-1}(x_{\pi(i)}), i = 1, 2, ..., \tilde{N}$$
        basic training for weak classifiers (parallel): Based on the decision tree (CART).
        calculating residuals (parallel): $\{R_{jkm}\}_{j=1}^{J} = J - terminal\_node\_tree(\{\tilde{y}_{\pi(i)k}, x_{\pi(i)}\}_1^{\tilde{N}})$
        calculating the optimal weak classifier coefficients :
            $$\rho_{jkm} = \arg\min \frac{C-1}{C}\frac{\sum_{x_{\pi(i)} \in R_{jkm}} \tilde{y}_{\pi(i)k}}{\sum_{x_{\pi(i)} \in R_{jkm}} |\tilde{y}_{\pi(i)k}| (1 - |\tilde{y}_{\pi(i)k}|)}, j = 1, 2, ..., J$$
        generating new weak classifiers :
            $$F_{km}(x) = F_{km-1}(x) + \sum_{j=1}^{J} \rho_{jkm} 1(x \in R_{jkm})$$
    end for
    training error detection:
    sampling : $\{test(m)\}_1^S = rand\_perm\,\{m\}_1^N$
    detection error (parallel): $err = predict(\{test(m)\}_1^S)$
    if $err < err_{\min}$ then
        exit from weak classifiers cycle
    end if
end for
obtaining a combination set of classification trees :
    $$\tilde{F}(x) = F_{MC}(x) = \sum_{m=1}^{M}\sum_{k=1}^{C} F_{km}(x)$$
```

Stochastic gradient boosting algorithm for its high accuracy rate received wide acclaim and is considered one of the most effective methods of statistical learning in classification, but its operational performance is poor, we propose two ways to improve its running performance: local parallelization and error detection shorten training times of weak classifiers.

The increasing popularity of multi-core processors to enhance the running performance of traditional algorithms provides another effective way. We propose a bottleneck module by way of parallel processing to enhance the stochastic gradient algorithm to improve running performance. First we consider parallel algorithms necessary and sufficient condition:

1) parallel algorithms have obvious advantages in large scale computing, and stochastic gradient boosting algorithm in step 2,5,6,7,13 involving the operation of the entire training samples, and general training samples exceeding thousands of pieces of data, so we consider parallel processing at these steps.

2) parallel algorithm must have a premise which is separability. Stochastic gradient boosting algorithm can not directly do paralleling at whole, because the algorithm is a additive model, each weak classifier training data is from error residuals of former process. So we can not be parallelized algorithm from the beginning of step 1,3. We can only do a local parallel processing.

Secondly, we tested the algorithm's main bottleneck module (refer to the module which has the inner loop in thousands) (see Table 1). The running count is the total count of overall algorithm (outer loop), The running time is running time of a single module running once. The running time of sampling and prediction is scale of microseconds, paralleling processing achieves few performance improvement. However calculation and residual gradient are in milliseconds. Parallel processing performance has significantly improved. The weak classifier training gain more performance improvement (10 milliseconds scale), since the whole number of cycles is up to 1000 times, thus improving overall training capability will reach 10 seconds. So parallel processing algorithms is necessary when algorithm involves huge data or are time-consuming.

In addition, the stochastic gradient algorithm to enhance the performance bottleneck lies in the number of basic classifiers. We tested the relationship between the number of classifiers and accuracy on the Optdigits data sets (see Figure 1). Classification accuracy was found to significantly increase with the increasement of the number of iteration process. The accuracy rate increase is not very obvious, even stagnation when iterations is up to 25. So, it is necessary to control the total number of iterations through the detection accuracy of the test sample in the training phase. To this end we introduce random sampling in order to optimize the training error detection methods to improve the stochastic gradient iterations through the step 11 to 16 (see Algorithm 1).

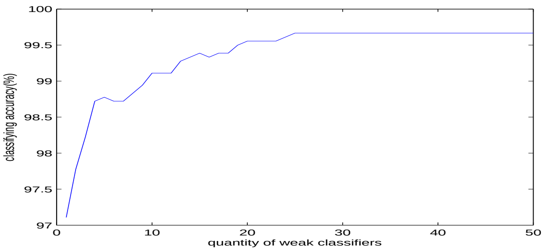


Figure 1: The relationship between classifying accuracy and quantity of weak classifiers on Optdigits data set.

3 Enhance image classification based on fast stochastic gradient boosting

This article discusses the general image classification methods and processes, we propose a fast stochastic gradient boosting to enhance image classification based the framework in Figure 2. First, we the extract image feature descriptor though PHOW features which are improved multi-scale dense SIFT descriptors, including basic steps: At first step, dense SIFT descriptors are calculated by dividing image into different scales with a fixed pixel Box (see Figure 3 line (a)). The descriptors of each grid point is calculated with four different diameter circular masks; For the second step, after extracting K-means clustering for the descriptors, a histogram is formed. At third step, we sum histogram pyramid to build space feature descriptors (see Figure 3 line (b)). Secondly, the characteristics of additive kernel transformation can generate better description of features. For finite dimensional distribution (histogram)

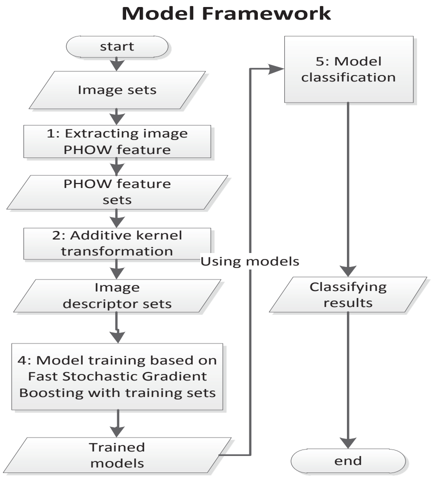

Model Framework
start
Image sets
1: Extracting image PHOW feature
PHOW feature sets
2: Additive kernel transformation
Image descriptor sets
4: Model training based on Fast Stochastic Gradient Boosting with training sets
Trained models
Using models
5: Model classification
Classifying results
end

Figure 2: The framework of image classification.

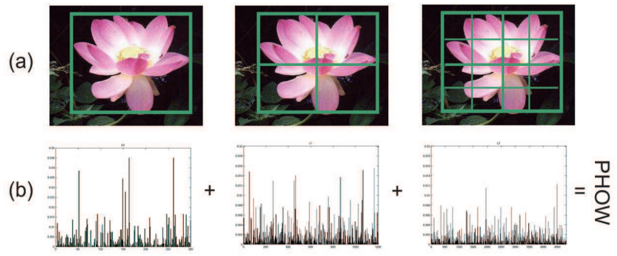


Figure 3: Image appearance representation based on PHOW.

x, y, Additive kernel is defined as :

$$K(x,y) = \sum_{b=1}^{B} k(x_b, y_b) \tag{1}$$

Here, b is a histogram of the number of each sub-grid, B is the total number of sub-grid, x_b, y_b is the distribution of every little grid, $k : \mathbb{R}_0^+ \times \mathbb{R}_0^+ \longrightarrow \mathbb{R}_0^+$ in the non-negative real number is a positive definite kernel. We proposed Vedaldi's χ^2 kernel transform for feature transformation.

Finally, We use the feature descriptors for fast stochastic gradient boosting algorithm to enhance the performance of classification model. At testing stage, we also need to extract features, then do kernel transformation of PHOW to form feature descriptor, again use a classification model to prediction. Our biggest advantage is that the entire framework is simple, good computing performance, and suitable for multi-category classification of natural images.

3.1 Experimental data sets

Optdigits data set[27] is a collection of data set standardized extracting of bit image by the U.S. National Institute of Standards and Technology handwritten Optical Character Recognition. It has 64 positive integers of feature information, the range is from 0 to 16. This data set consists of 5620 instances, belonging to 10 categories. we randomly selected 10%, 20% and 30%, 40% and 50% of total sample as the training sample, and the rest for test samples.

15-Scenes data set[28] is processed in accordance with the flow chart of our proposed framework (see Figure 2). The original 15-Scenes data set consists of 15 data categories, a total of 4485 images. We randomly select 10% and 20%, 30% and 40% and 50% of each class sample for the training sample, and the rest used to do the test samples. We use the PHOW descriptor for image features to describe each image. After kernel transformation with additive eventually, we get 36,000 dimensional feature descriptors for a single image.

3.2 Parameters

1) Maximum decision tree depth: The default value was set to 1. With the value increase, classification accuracy and running time will increase. We set maximum depth to 2, 4, 6, 8, 10, 12, 14, 16, 18 and 20 respectively. We found the highest accuracy rate when the maximum depth is 10. With similar way, we found that entropy rules of split consideration criterion get better performance.

2) The maximum depth of random random forest was similar with decision tree. The number of decision trees: we tested the value of 20, 40, 60, 80, 100, 120, 140, 160, 180 and 200 respectively. We found that the number increases, the execution time also increases, and after over the value of 100, the improvement of the accuracy rate was not obvious. So we set it to 100. The accuracy of the random forest was used to control the iteration. We tested the value of 0.0001, 0.001, 0.01 and 0.1 respectively. We found that the smaller the value was, the longer the execution time was, and after over the value of 0.001, the accuracy had no substantially change. We set it to 0.001.

3) The extreme random tree's settings was similar with random forests for consistent comparing standard.

4) With similar ways, we found that we get better performance (good balance in accuracy and running time) when the stochastic gradient enhance maximum tree depth is set to 10, cross entropy loss chosen for loss function type , 0.1 set to shrinkage factor, 0.8 set to proportional sampling under, and 100 chosen for maximum lift.

5) Similarly, in order to enhance the fast stochastic gradient boosting based on the stochastic gradient boosting, the control error was set to: 0.0001, random verification sampling ratio was set as follows: 50%.

3.3 Experimental setup

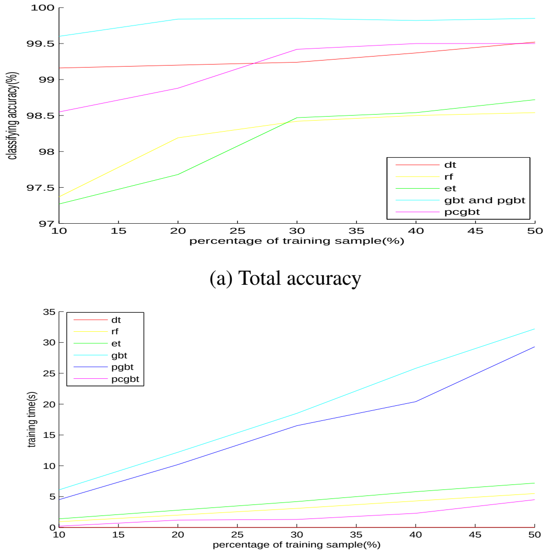
We used C++ of Microsoft visual studio 2012 to program with opencv2.4.3[24], Intel TBB[25] and darwin 1.6 platform[26]. We tested results in win7 (64) platform with hardware of Intel P6100 dual-core CPU and 6GB memory.

We converted the data set to comma delimited $M \times N$ in the form of a text file, M represented the number of (ie, the number of records) data rows, N was the number of each data attribute values. The final column was class marker.

We used TBB parallel libary for parallel processing. In Algorithm 1: At steps of 2,5,7,13, We used TBB with $tbb :: parallel\_for$. At step 6 involving recursive tree, we used TBB with $tbb :: task\_list$ to achieve parallelism.

In order to better reflect the effectiveness of the proposed framework, we have two types of data sets in the test, Optdigits for low-dimensional data, 15-Scenes for high-dimensional data, and extract the sample test to verify the practicality for different circumstance.

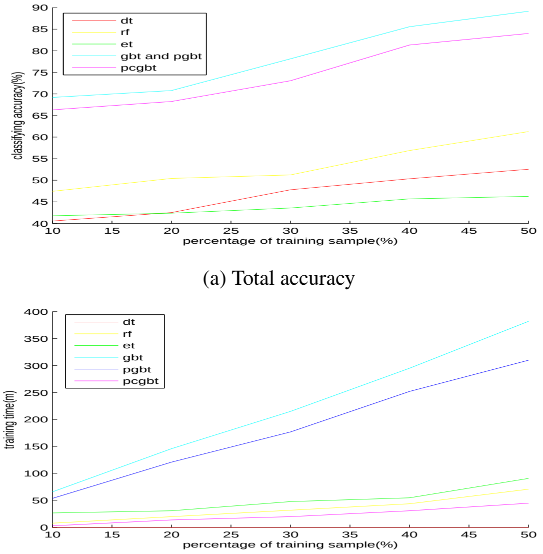
In addition to verify the feasibility of stochastic gradient boosting algorithm and its fast versions, our experiments compared the decision tree, random forest, extreme random tree, stochastic gradient boosting , stochastic gradient boosting with error check and SVM (support vector machine). Experiment results were indicated in Table 2, 3, and Figure 4 to 7.



(a) Total accuracy

(b) Training time

Figure 4: The performances of six algorithms on Optdigits.



(a) Total accuracy

(b) Training time

Figure 5: The performances of six algorithms on 15-Scenes.

3.4 Experimental results and analysis

Table 2-3 is a averaged results tested three times under the same conditions. Where h is hours, m represents minutes, s is seconds, ms is milliseconds, such as: 1h2m3s4ms represents 1 hour, 2 minutes, 3 seconds, and 4 milliseconds. dt is a decision tree, rt is random forests, et is extreme random tree, gbt is stochastic gradient boosting, pgbt is fast stochastic gradient boosting, pcgbt is the fast stochastic gradient boosting with error check, and SVM is support vector machine.

1) With increase of the proportion of each sampling, algorithm accuracy rate increases, however the corresponding training time also increases. This indicates the adequacy of the training sample for classification accuracy is critical, but the running performance will be affected in the training. In practical applications, we should consider the two factors, and try to find the best balance between them.

2) Total accuracy comparison: As can be seen from Table 2-3 and Figure 4-5 (a), stochastic gradient boosting and fast stochastic gradient boosting have same overall accuracy. Overall accuracy on 15-Scenes data set from high to low is stochastic gradient boosting, stochastic gradient boosting with error detection, random forest, extreme random tree and decision tree. The main difference on Optdigits data set lies in that decision tree was significantly better than random forests, furthermore compared with random gradient boosting, the accuracy of our stochastic gradient boosting with error check also has a certain decline of accuracy, but is still significantly better than the decision tree and random forest.

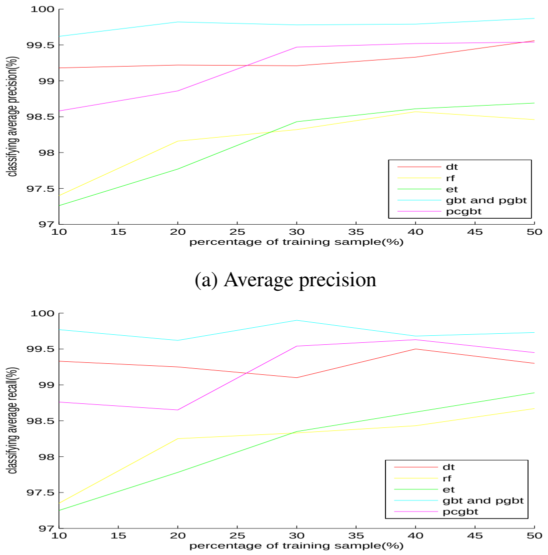
3) Running time comparison: From Table 2 to 3 and Figure 4 to 5 (b) shows that the training runtime performance in descending order is decision tree, fast stochastic gradient boosting with error detection, random forests and random forests extreme, fast stochastic gradient boosting and stochastic gradient boosting. Stochastic gradient boosting with error detection is about 10 times fast than the original stochastic gradient boosting on the data set Optdigits and 8 times on 15-Scenes data set.

4) Average recall,average precision and total accuracy comparisons: from figure 4 to 7, we can see that the cures have similar curve tendency. This shows that total accuracy basically reflect the performance of classifier on Optdigits and 15-Scenes data set respectively.

5) Comparison with support vector machine: from table 2 and table 3, we can see that the total accuracy of SVM is superior to decision tree, random forest trees and extremely random trees, however inferior to stochastic boosting tree based methods. Furthermore, on 15-Scenes data set SVM is failed when the training sampling percentages reach 40% and 50%. On the side of the training time, SVM is slower than decision tree, random forest trees and extremely random trees, but faster the stochastic boosting tree based methods.

4 Conclusions

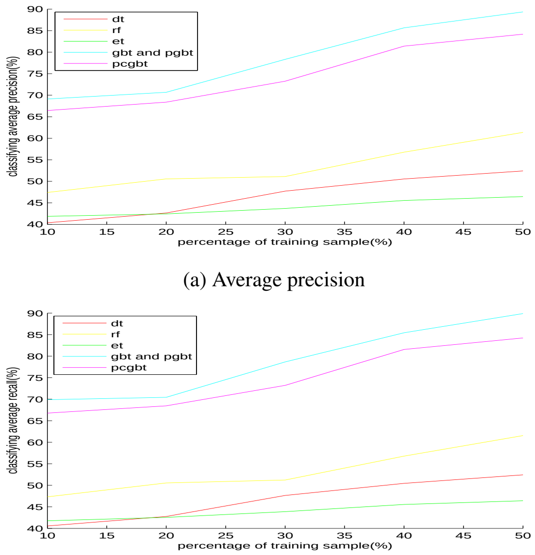
By comparing each algorithm, we have following the experimental findings: 1) Stochastic gradient boosting is significantly better than decision tree, random forest and extreme random tree. 2) We proposed parallel stochastic gradient boosting algorithm to enhance the running performance. Experimental result of our method is significantly better than the original stochastic gradient boosting algo-



(a) Average precision

(b) Average recall

Figure 6: The performances of six algorithms on Optdigits.



(a) Average precision

(b) Average recall

Figure 7: The performances of six algorithms on 15-Scenes.

rithm. Furthermore, fast stochastic gradient boosting with error detection improves running performance to a new stage, while keeping the overall accuracy comparing to the original stochastic gradient boosting. Experiments testify that our improvement ways are effective and practical.

The main contributions of this paper are :

1) We presents a framework based on PHOW features and fast stochastic gradient boosting for natural image classification. From training sample selection, feature extracting, classifier selection to the last performance evaluation, we give a detailed analysis and commentary.

2) We analyze the running performance and bottlenecks of the stochastic gradient boosting. According to the circumstance of bottlenecks and current widely used multicore computing, we presented modified stochastic gradient boosting to improve the performance of by local parallelism.

3) Due to reason of that increasing number of weak classifiers does not always bring better accuracy, and to further reduce the space and time of weak classifier training iterations, we introduce an error control mechanism in training phase to reduce the number of iterations of the method at the expense of a certain degree of accuracy degeneration. However, by this way we get further improvement of the running performance.

4) In this paper, a parallel realization of serial algorithm are thoroughly discussed. Taking stochastic gradient boosting as example, we proposed a well-established ideas and methods of these kind of problems, namely: 1) to detect bottlenecks, determining the optimization core; 2) the task segmentation, transforming a serial program by parallelization ideas; 3) implementation based on TBB parallel architecture.

References

[1] R. M. Haralick, K. Shanmugam, I. Dinstein. Textural features for image classification.*IEEE Transactions on Systems, Man and Cybernetics*,(1973), SMC-3(6):610−621

[2] T. Ojala, M. Pietikainen, T. Maenpaa. Multiresolution gray-scale and rotation invariant texture classification with local binary patterns.*IEEE Transactions on Pattern Analysis and Machine Intelligence* , (2002), 24(7): 971−87

[3] D. G. Lowe. Distinctive image features from scale-invariant keypoints.*International journal of computer vision*, (2004), 60(2): 91−110

[4] N. DALAL, B. TRIGGS. Histograms of oriented gradients for human detection. In: Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition, New York, USA: IEEE, (2005). 886−893.

[5] T. Leung, J. Malik. Representing and recognizing the visual appearance of materials using three-dimensional textons.*International journal of computer vision*, (2001), 43(1): 29−44.

[6] L. Nanni, A. Lumini. Heterogeneous bag of features for object/scene recognition.*Applied Soft Computing*, (2013), 13(4): 2171−2178.

[7] A. Sinha, S. Banerji, C. Liu. Novel color Gabor-LBP-PHOG (GLP) descriptors for object and scene image classification.In: Proceedings of the Eighth In-

dian Conference on Computer Vision, Graphics and Image Processing, Mubai, (2012): 581−588.

[8] A. Bosch, A. Zisserman, X. Muoz. Image classification using random forests and ferns.In: Proceedings of the International Conference on Computer Vision, Rio de Janeiro, (2007):1−8.

[9] A. Vedaldi, A. Zisserman. Efficient additive kernels via explicit feature maps.*IEEE Transactions on Pattern Analysis and Machine Intelligence*, (2012), 34(3): 480−492.

[10] P. Kamavisdar, S. Saluja, S. Agrawal. A Survey on Image Classification Approaches and Techniques.In: International Journal of Advanced Research in Computer and Communication Engineering, (2012), 2(1):1005−1009.

[11] R. M. Haralick, K. Shanmugam, I. H. Dinstein. Textural features for image classification.*IEEE Transactions on Systems, Man and Cybernetics*, (1973), SMC-3(6): 610−621.

[12] G. Ridgeway. Generalized Boosted Models: A guide to the gbm package [Online], available: https://code.google.com/p/ gradientboostedmodels/.

[13] O. Chapelle, P. Haffner, V. N Vapnik. Support vector machines for histogram-based image classification.*IEEE Transactions on Neural Networks*, (1999), 10(5): 1055−1064.

[14] G. M. Foody, A. Mathur. A relative evaluation of multiclass image classification by support vector machines. *IEEE Transactions on Geoscience and Remote Sensing*, (2004), 42(6): 1335−1343.

[15] L. Breiman, J. Friedman, C. J. Stone, et al. *Classification and regression trees.*New York: Chapman & Hall/CRC, (1984).

[16] R. Lawrence, A. Bunn, S. Powell, et al. Classification of remotely sensed imagery using stochastic gradient boosting as a refinement of classification tree analysis.*Remote sensing of environment*, (2004), 90(3): 331−336.

[17] L. Breiman. Bagging predictors. *Machine learning*, (1996), 24(2): 123−40

[18] L. BREIMAN. Random forests. *Machine learning*, (2001), 45(1): 5−32.

[19] P. GEURTS, D. ERNST, L. WEHENKEL. Extremely randomized trees. *Machine learning*, (2006), 63(1): 3−42.

[20] E. Bauer, R. Kohavi. An empirical comparison of voting classification algorithms: Bagging, boosting, and variants. *Machine learning*, (1998), 36(1): 105−139.

[21] Y. Freund, R. E. Schapire. Experiments with a new boosting algorithm. In: Proceedings of International Conference on Machine Learning,Bari.(1996) :148−156.

[22] K. P. Murphy. *Machine Learning: a Probabilistic Perspective.* Massachusetts: the MIT press. (2012):553−562

[23] J. H. Friedman. Stochastic gradient boosting. *Computational Statistics and Data Analysis*, (2002), 38(4): 367−378.

[24] D. Abram, T. Pribanic, H. Dzapo, M. Cifrek. A brief introduction to OpenCV.In: Proceedings of the 35th International Convention, Opatija,(2012). 1725−1730.

[25] J. Reinders. *Intel threading building blocks: outfitting C++ for multi-core processor parallelism.* Gravenstein: O'Reilly Media, Inc. (2010).

[26] S. Gould. DARWIN: A Framework for Machine Learning and Computer Vision Research and Development.*Journal of Machine Learning Research*, (2012). 13(12): 3499−3503.

[27] K. Bache, Lichman. UCI machine learning repository[Online]. http://archive.ics.uci.edu/ml.(2013).

[28] S. Lazebnik, C. Schmid, Ponce J. Beyond bags of features: Spatial pyramid matching for recognizing natural scene categories. In: Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition, New York, USA: IEEE, (2006). 2169−2178.

Modules	Sampling	calculating the gradient	weak classifier training	computing residuals	Prediction
Running count	100	1000	1000	1000	100
Serial Time	$170us$	$2.2ms$	$62ms$	$1.6ms$	$0.5ms$
Parallel Time	$150us$	$1.3ms$	$50ms$	$1.2ms$	$0.4ms$

Table 1: Serial and parallel executing time of bottleneck modules on Optdigits data set.

Accuracy (%) / Time	10 %	20 %	30 %	40 %	50 %
dt	$99.16/6.0ms$	$99.20/17.0ms$	$99.24/19.9ms$	$99.37/25.0ms$	$99.52/34.0ms$
rt	$97.37/933ms$	$98.19/2.0s$	$98.42/3.1s$	$98.50/4.3s$	$98.54/5.5s$
ert	$97.27/1.4s$	$97.68/2.8s$	$98.47/4.2s$	$98.54/5.8s$	$98.72/7.2s$
gbt	$99.60/6.1s$	$99.84/12.2s$	$99.85/18.5s$	$99.82/25.8s$	$99.85/33.2s$
pgbt	$99.60/4.5s$	$99.84/10.2s$	$99.85/16.5s$	$99.82/20.4s$	$99.85/29.3$
pcgbt	$98.55/206ms$	$98.88/1.2s$	$99.42/1.3s$	$99.50/2.3s$	$99.50/4.5s$
SVM	$98.52/406ms$	$98.60/1.8s$	$98.62/2.1s$	$98.73/3.5s$	$98.80/6.5s$

Table 2: The Comparison of accuracy and running time of six algorithms on Optdigits with different sampling.

Accuracy (%) / Time	10 %	20 %	30 %	40 %	50 %
dt	$40.56/3s$	$42.53/6s$	$47.80/10s$	$50.34/14s$	$52.56/16s$
rt	$47.44/8m$	$50.43/20m$	$51.24/32m$	$56.89/44m$	$61.28/1h11m$
ert	$41.77/27m$	$42.37/31m$	$43.58/48m$	$45.68/55m$	$46.27/1h31m$
gbt	$69.18/1h6m$	$70.77/2h26m$	$78.13/3h35m$	$85.56/4h55m$	$89.15/6h22m$
pgbt	$69.18/54m$	$70.77/2h1m$	$78.13/2h/57m$	$85.56/4h12m$	$89.15/5h10m$
pcgbt	$66.32/3m$	$68.25/14m$	$73.05/20m$	$81.33/31m$	$84.01/45m$
pcgbt	$66.32/3m$	$68.25/14m$	$73.05/20m$	$81.33/31m$	$84.01/45m$
SVM	$65.44/10m$	$67.31/34m$	$72.16/66m$	$invalid$	$invalid$

Table 3: The comparison of accuracy and running time of six algorithms on 15-Scenes with different sampling.

Word Sense Disambiguation Using an Evolutionary Approach

Mohamed El Bachir Menai
Department of Computer Science, College of Computer and Information Sciences
King Saud University, P.O.Box 51178, Riyadh 11543, Saudi Arabia
menai@ksu.edu.sa
http://faculty.ksu.edu.sa/menai

Keywords: evolutionary algorithms, genetic algorithms, natural language understanding, word sense disambiguation

Received: July 22, 2013

Word sense disambiguation is a combinatorial problem consisting in the computational assignment of a meaning to a word according to a particular context in which it occurs. Many natural language processing applications, such as machine translation, information retrieval, and information extraction, require this task which occurs at the semantic level. Evolutionary computation approaches can be effective to solve this problem since they have been successfully used for many NP-hard optimization problems. In this paper, we investigate main existing methods for the word sense disambiguation problem, propose a genetic algorithm to solve it, and apply it to Modern Standard Arabic. We evaluated its performance on a large corpus and compared it against those of some rival algorithms. The genetic algorithm exhibited more precise prediction results.

Povzetek: Razločitev pomena besed je v tem prispevku izpeljana z evolucijskim pristopom.

1 Introduction

Ambiguity is a key feature of natural languages. That is, words can have different meanings (polysemy), depending on the context in which they occur. Humans deal with language ambiguities by acquiring and enriching common sense knowledge during their lives. However, solving computationally the ambiguity of words is a challenging task, since it relies on knowledge, its representation, extraction, and analysis. In Arabic language, ambiguity is present at many levels [30], such as homograph, internal word structure, syntactic, semantic, constituent boundary, and anaphoric ambiguity. The average number of ambiguities for a token in Modern Standard Arabic (MSA) is 19.2, whereas it is 2.3 in most languages [30].

Word sense disambiguation (WSD) is a challenging task in the area of natural language processing (NLP). It refers to the task that automatically assigns the appropriate sense, selected from a set of pre-defined senses for a polysemous word, according to a particular context. Indeed, the identification of one word sense is related to the identification of neighboring word senses. WSD is necessary for many NLP applications and is believed to be helpful in improving their performance such as machine translation, information retrieval, information extraction, part of speech tagging, and text categorization. WSD has been described as an AI-complete (Artificial Intelligence-complete) problem [53] that is analogous to NP-complete problems in complexity theory. It can be formulated as a search problem and solved approximately by exploring the solution search space using heuristic and meta-heuristic algorithms. Several approaches have been investigated for WSD in occidental languages (English, French, German, etc.), including knowledge-based approaches and machine learning-based approaches. However, research on WSD in Arabic language is relatively limited [5,6,17,24–27,38].

Evolutionary algorithms (EAs) are search and optimization methods inspired by biological evolution: natural selection and survival of the fittest in the biological world. Several types of EAs were developed, including genetic algorithms (GAs) [41], evolutionary programming (EP) [32], evolution strategies (ES) [69,76] and genetic programming (GP) [45]. EAs are among the most popular and robust optimization methods used to solve hard optimization and machine learning problems. They have been widely and successfully applied in several real world applications [55] and research domains. These include NLP research, such as query translation [20], inference of context free grammars [43], tagging [8], parsing [7], and WSD [22,35,84]. Araujo [9] has written a survey paper on how EAs are applied to statistical NLP, which is highly recommended.

In this paper, we study the potential of GAs in formulating and solving the WSD problem, apply them to MSA, and compare them with some existing methods. We implemented and experimented different variants of GAWSD (GA for Arabic WSD) resulting in the introduction of a competitive approach for WSD. The rest of the paper is organized as follows. The next section presents a brief overview of EAs. Section 3 contains a brief introduction to WSD, and presents the main approaches to solve it. Section 4 describes Arabic language peculiarities and challenges. Section 5 presents the proposed approach to WSD, and describes in detail the proposed algorithm. Section 6 reports the test results, and Section 7 discusses them. Finally, Section 8 concludes this paper and emphasizes some future directions.

2 Evolutionary algorithms

EAs are built around four key concepts [21]: population(s) of individuals competing for limited resources, dynamic changing populations, suitability of an individual to reproduce and survive, and variational inheritance through variation operators.

EAs are categorized as "generate and test" algorithms that involve growth or development in a population of chromosomes in genotype space (individuals containing genes) of candidate solutions in phenotype space (real features of an individual). An evaluation function called the *fitness function*, defined from chromosome representation, measures how effective the candidate solutions are as a solution to the problem. Variation operators such as *recombination* (or *crossover* in case of recombination of two parents) and *mutation* are applied to modify the individual content and promote diversity.

Algorithm 1: Evolutionary Algorithm

```
Initialize $P(1)$;
$t \leftarrow 1$;
while not exit criterion do
    evaluate $P(t)$;
    selection;
    recombination;
    mutation;
    survive;
    $t \leftarrow t + 1$;
```

The basic steps of an EA are outlined in Algorithm 1. An *initial population*, $P(1)$, is randomly generated and a selection process (*selection*) is then performed to select parents based on the fitness of individuals (*evaluate*). The *recombination* and *mutation* operators are applied on parents to obtain a population of offspring. The population is renewed (*survive*) by selecting individuals from the current population and offspring for next generation $(t + 1)$. This evolutionary process continues until a termination condition, *exit criterion*, is reached.

GAs [41] are the most traditional EAs which are based on biological genetics, natural selection, and emergent adaptive behavior. They are associated to the use of binary, integer, or real valued vectors for the chromosome representation. The crossover and mutation are the genetic operators. The crossover is the main operator (applied with a high probability), and the mutation is the secondary one (applied with a low probability). The main steps of a GA are outlined in Algorithm 2 [15]. GP [45] can be considered as an extension of GAs in which each individual is a computer program represented by a rooted tree. In this case, the fitness function determines how well a program is able to solve the problem.

Algorithm 2: Genetic algorithm

```
input  : $Population_{size}$, $Problem_{size}$, $P_{crossover}$,
         $P_{mutation}$
output: $S_{best}$
$Population \leftarrow$ Initialize($Population_{size}$,
$Problem_{size}$);
Evaluate($Population$);
$S_{best} \leftarrow$ BestSolution($Population$);
while not exit criterion do
    $Parents \leftarrow$ SelectParents ($Population$);
    $Children \leftarrow \phi$;
    for $Parent_1, Parent_2 \in Parents$ do
        $(Child_1, Child_2) \leftarrow$ Crossover
        $(Parent_1, Parent_2, P_{crossover})$;
        $Children \leftarrow$ Mutate $(Child_1, P_{mutation})$;
        $Children \leftarrow$ Mutate $(Child_2, P_{mutation})$;
    Evaluate($Children$);
    $S_{best} \leftarrow$ BestSolution($Children$);
    $Population \leftarrow$ SelectToSurvive
    ($Population, Children$);
Return($S_{best}$)
```

3 Classification methods for word sense disambiguation

WSD can be described as the task of assigning the appropriate sense to all or some of the words in the text. More formally, given a text T as a sequence of words or bag of words $\{w_1, w_2, \cdots, w_k\}$, the WSD problem asks to identify a mapping A from words w_i to senses $Senses_D(w_i)$ encoded in a dictionary D. $A(w(i))$ is the subset of the senses of w_i which are appropriate in the context T [62].

A WSD system includes mainly four elements: word senses selection, external use of knowledge resources, context representation, and selection of automatic classification method. The first element, selection of word senses, is concerned with the sense distinction (sense inventory) of a given word. The second element, external knowledge sources, involves a repository of data consisting of words with their senses. Two main kinds of resources are distinguished: structured resources and unstructured resources. The third element of WSD is concerned with the representation of the context that aims to convert unstructured input text into a structured format to become suitable for automatic methods. The last element of WSD is the choice of the classification method. The key distinction between classification methods depends on the amount of knowledge and supervision quantified into them.

In the following, we survey the main classification methods used for WSD, as they represent a key issue in designing a WSD system.

Classification methods can be achieved using different approaches [62]: knowledge-based and machine learning-based approaches. Knowledge-based methods rely on external lexical resources, such as dictionaries and thesauri, whereas machine learning methods (supervised, unsupervised, or semi-supervised methods) rely on annotated or unannotated corpus evidence and statistical models. Other methods use both corpus evidence and semantic relations. They can be further categorized as token-based or type-based approaches. While token-based approaches associate a specific meaning with each occurrence of a word depending on the context in which it appears, type-based disambiguation is based on the assumption that a word is consensually referred with the same sense within a single text [62].

Other approaches have been considered, such as word sense dominance-based methods [46,54,59], domain-driven disambiguation [37], and WSD from cross-lingual evidence [33].

3.1 Knowledge-based methods for word sense disambiguation

Several knowledge-based methods for WSD have been proposed, including gloss-based methods, selectional preferences-based methods, and structural methods.

Gloss based-methods consist in calculating the overlap of sense definitions of two or more target words using a dictionary. Such methods include the well-known Lesk algorithm [50] and one of its variants proposed by Banerjee and Pedersen [11].

Selectional preferences (or restriction) based methods exploit association provided by word-to-word, word-to-class, or class-to-class relations to restrict the meaning of a word occurring in a context, through grammatical relations [62]. Several techniques have been proposed to model selectional preferences, such as selectional associations [70,72], tree cut models [51], hidden Markov models [1], class-based probability [2,19], and Bayesian networks [18]. An application of such associations to expanding an Arabic query of a search engine [5] shows that the performance of the system can be increased by adding more specific synonyms to the polysemous terms.

Structural approaches are semantic similarity-based methods and graph-based methods. The main idea behind these approaches is to exploit the structure of semantic networks in computational lexicons like WordNet [31], by using different measures of semantic similarity. Some examples of knowledge-based systems include Degree [61] and Personalized PageRank [3].

Similarity-based methods are applicable to a local context, whereas graph-based methods are applicable to a global context. Similarity-based methods select a target word sense in a given context based on various measures of semantic similarity, such as those introduced by Rada et al. [68], Sussna [77], Leacock and Chodorow [47], Resnik [71], Jiang and Conrath [42], and Lin [52]. Elghamry [27] proposed coordination-based semantic similarity for disambiguating polysemous and homograph nouns in Arabic, based on the assumption that nouns coordinating with an ambiguous noun provide bootstraps for disambiguation.

Graph-based methods select the most appropriate sense for words in a global context using lexical chains (sequence of semantically related words by lexicosemantic relations) [62]. Many computational models of lexical chains have been proposed, including those of Hirst and St-Onge [40], Galley and McKeown [34], Harabagiu et al. [39], Mihalcea et al. [57], and Navigli and Velardi [63].

3.2 Machine leaning methods for word sense disambiguation

There are three classes of machine learning methods: supervised, unsupervised, and semi-supervised methods. All of them have been largely applied to WSD.

The most popular supervised WSD methods include decision lists [82], decision trees [60], naïve Bayes classifiers [66], artificial neural networks [60,79], support vector machines [29,48,85], and ensemble methods [28]. Farag and Nürnberger [6] used a naïve Bayes classifier to find the correct sense for Arabic-English query translation terms by using bilingual corpus and statistical co-occurrence.

Unsupervised WSD methods usually select the sense from the text by clustering word occurrences. Through measuring the similar neighboring words, new occurrences can be classified into clusters/senses. Since unsupervised methods do not use any structured resource, their assessment is usually difficult. The main approaches to unsupervised WSD are context clustering [67,75], word clustering [14,65], and co-ocurrence graphs [80]. Diab [25] introduced an unsupervised method called SALAAM (stands for Sense Annotations Leveraging Alignments And Multilinguality) to annotate Arabic words with their senses from an English WordNet using parallel Arabic-English corpus based on translational correspondences between Arabic and English words. Lefever et al. [49] used a multilingual WSD system, called ParaSense, where the word senses are derived automatically from word alignments on a parallel corpus.

To address the lack of the training data problem, semi-supervised WSD methods use both annotated and unannotated data to build a classifier. The main semi-supervised WSD methods are based on a bootstrapping process which starts with a small amount of annotated data (called seed data) for each word, a large corpus of unannotated data, and one or more classifiers. The seed data are used to train the classifier using a supervised method. This classifier then uses the unannotated data to increase the amount of annotated data and decrease the amount of unannotated data. This process is repeated until achieving an amount threshold of unannotated data. Co-training and self-training are two bootstrapping approaches used in WSD. Co-training uses two classifiers for local and global information (e.g. [56]). Self-training uses only one classifier that merges the two types of information. An example of self-training approach is illustrated by Yarowsky algorithm [83].

3.3 Evolutionary algorithms for word sense disambiguation

Gelbukh et al. [35] used a GA as a global optimization method (the total word relatedness is optimized globally) to tackle WSD problem. An individual is represented by a sequence of natural numbers of possible word senses retrieved from a dictionary, and the Lesk measure [50] is used to evaluate its fitness. The experimental results obtained on Spanish words show that this method gives better results than existing methods which optimize each word independently.

Decadt et al. [22] used a GA to improve the performance of GAMBL, a WSD system. WSD is formulated as classification task distributed over word experts. A memory-based learning method is used to assign the appropriate sense to an ambiguous word, given its context. The feature selection and algorithm parameter optimization are performed jointly using a GA. The experimental results obtained on Senseval-3 English all-words task, show the constructive contribution of the GA on system performance with a mean accuracy of 65.2%.

Zhang et al. [84] proposed a genetic word sense disambiguation (GWSD) algorithm to maximize the semantic similarity of a set of words. An individual is represented by a sequence of natural numbers of possible word senses retrieved from WordNet. The length of the chromosome is the number of words that need to be disambiguated. The fitness function used is based on the Wu-Palmer similarity measure [81] in which the domain information and the frequency of a given word sense are included. The evaluation of the algorithm gives a mean recall of 71.96%.

4 Arabic language

4.1 Arabic language characteristics

Arabic language belongs to the Afro-Asian language group. Its writing is right to left, cursive, and does not include capitalization. Arabic letters change shape according to their position in the word, and can be elongated by using a special dash between two letters.

The language is highly inflectional. An Arabic word may be composed of a stem plus affixes (to refer to tense, gender, and/or number) and clitics (that include some prepositions, conjunctions, determiners, and pronouns). Words are obtained by adding affixes to stems which are in turn obtained by adding affixes to roots.

Diacritization or vocalization in Arabic, consists in adding a symbol (a diacritic) above or below letters to indicate the proper pronunciation and meaning of a word. The absence of the diacritization in most of Arabic electronic and printed media poses a real challenge for Arabic language understanding. Arabic is a pro-drop language: it allows subject pronouns to drop, like in Italian, Spanish, Chinese, and Japanese [30].

Dealing with ambiguity in Arabic is considered as the most challenging task in Arabic NLP. There are two main levels of Arabic ambiguity [10,30]: (1) Homographs are words that have the same spelling, but different meanings. The main cause of homographs is due to the fact that the majority of digital documents do not include diacritics; (2) Polysemy is the association of one word with more than one meaning. Ambiguity in Arabic can be also present in other levels, such as: internal word structure ambiguity, syntactic ambiguity, semantic ambiguity, constituent boundary ambiguity, and anaphoric ambiguity [30].

MSA is the subject of this research. It is the language of modern writing and formal speaking. It is the language universally understood by Arabic speakers around the world. In contrast, Classical Arabic (CA) is the language of religious teaching, poetry, and scholarly literature. MSA is a direct descendent of CA [12].

4.2 Arabic text preprocessing

Text preprocessing consists in converting a raw text file into a well-defined sequence of linguistically-meaningful units, such as characters, words, and sentences [64]. It includes the following tasks:

- Tokenization or sentence segmentation is the process of splitting the text into words.
- Stop-word removal is the process of filtering a text from the stop-words, such as prepositions and punctuation marks, assuming that they do not deeply alter the meaning of the text.
- Stemming is the process of removing prefixes and suffixes to extract stems.
- Rooting is the process of reducing words to their roots.

There are some well-known algorithms for morphological analysis, such as Khoja's stemmer [44], Buckwalter's morphological analyzer [16], the Tri-literal root extraction algorithm [4], MADA (Morphological Analysis and Disambiguation for Arabic) [38,73], and AMIRA [23].

5 Proposed approach

Amongst the various methods presented in Section 3, the mostly used methods for WSD are supervised WSD and knowledge-based methods. Supervised WSD methods achieve better performance than knowledge-based methods given large training corpora, but they are generally limited to small contexts. Knowledge-based methods can exploit all available knowledge resources, such as dictionaries and thesauri, but they require exponential computational time as the number of words increases. Our approach consists in approximating solutions to WSD problem by using GAs to improve the performance of a gloss-based method. We adopt a similar individual (or chromosome) representation to the one presented in [8,35], but different evaluation functions of the individual fitness and different selection methods. To the best of our knowledge, there is no published research proposing an evolutionary computing-based approach to solve the WSD problem in Arabic language.

Algorithm 3 outlines the main steps of the genetic algorithm for Arabic WSD (GAWSD).

A text T is transformed into a bag of words $\{w_1, w_2, \cdots, w_k\}$ in a preprocessing phase, including stop-word removal, tokenization, and rooting. The accuracy of Arabic WSD algorithms can be increased by reducing the words to their root form. A morphological analysis is then needed to extract the root form of the word. The comparative evaluation of Arabic language morphological analyzers and stemmers [74], namely Khoja's stemmer, the tri-literal root extraction algorithm, and the Buckwalter morphological analyzer, shows that Khoja's stemmer achieves the highest accuracy. In our algorithm, Khoja's stemmer is used to reduce words to their roots. The senses $Senses_{AWN}(w_i)$ of each word w_i are retrieved from Arabic WordNet (AWN) [13] as word definitions which are reduced in turn to bags of words. An GA is used to find the most appropriate mapping from words w_i to senses $Senses_{AWN}(w_i)$ in the context T. The best individual S_{best} returned by the GA, is decoded into the phenotype space to obtain the appropriate sense of words $WordsSense_{best}$.

Algorithm 3: GAWSD

```
input  : $T$, $k$, $Population_{size}$, $P_{crossover}$, $P_{mutation}$
output: $WordsSense_{best}$
$\{w_1, w_2, \cdots, w_k\} \leftarrow$ Preprocessing($T$);
for $i = 1, k$ do
    $Definitions(w_i) \leftarrow$ AWN($w_i$);
    $Senses_{AWN}(w_i) \leftarrow$
    Preprocessing($Definitions(w_i)$);
$S_{best} \leftarrow$ GA($Population_{size}$, $k$,
$Senses_{AWN}(w_i)_{i=1,k}$, $P_{crossover}$, $P_{mutation}$);
$WordsSense_{best} \leftarrow$ Decode($S_{best}$);
Return($WordsSense_{best}$)
```

To formulate the WSD problem in terms of GA, we need to define the following elements:

- A representation of an individual of the population.
- A method to generate an initial population.
- An evaluation function to determine the fitness of an individual.
- A description of the genetic operators (crossover and mutation).
- Methods to select parents for the mating pool and individuals to survive to the next generation.
- Values for the several algorithm parameters (population size, crossover and mutation rates, termination condition, tournament size, etc.).

More specifically, we propose the following formulation to solve the WSD problem. Alternative solutions for key elements of the algorithm, such as generation of initial population, fitness function, etc., will be considered to find out the appropriate resolution.

- An individual Ind_p represents a possible sequence of sense indexes assigned to the words in the context T. It is represented by a fixed-length integer string $Ind_p = \{SI^l(w_1), SI^m(w_2), \cdots, SI^r(w_k)\}$, where each gene $SI^j(w_i)$ is an index to one of possible senses of the word w_i: $SI^0(w_i), SI^1(w_i) \cdots SI^l(w_i) \cdots$.
- The initial population is generated according to one of the following schemes:
  – Random generation: The value of each gene of an individual is selected randomly from 1 to $SenseNum$ using the uniform distribution, where $SenseNum$ is the number of possible senses for the corresponding word.
  – Constructive generation: All the senses of a given word are distributed in a round-robin way to the corresponding gene of individuals in the population.
- The fitness function is measured by the word sense relatedness. Two different measures are considered: the Lesk measure [50] and one of its variants, called the extended Lesk measure. The Lesk measure calculates the sense which leads to the highest overlap between the sense definitions of two or more words. Formally, given two words w_1 and w_2, and their respective senses $Senses_{AWN}(w_1)$ and $Senses_{AWN}(w_2)$, for each two senses $S_1 \in Senses_{AWN}(w_1)$ and $S_2 \in Senses_{AWN}(w_2)$, the Lesk measure is defined by Equation 1,

$$score_{Lesk}(S_1, S_2) = |gloss(S_1) \cap gloss(S_2)| \quad (1)$$

where $gloss(S_i)$ represents the bag of words corresponding to the definitions of the sense S_i.
The extended Lesk measure calculates the overlap between the sense definitions of a target word and the words in its context. Formally, given a word w and its $context(w)$, for each sense S_i of w, the extended Lesk measure is defined as by Equation 2,

$$score_{extendedLesk}(S_i) = |context(w) \cap gloss(S_i)| \quad (2)$$

where $gloss(S_i)$ represents the bag of words corresponding to the definitions of the sense S_i.

- A single-point crossover operator combines two individuals (parents) to generate two new ones (children or offspring). The crossover point is chosen randomly.

- A single-point mutation operator creates a new individual by randomly changing the value of a selected gene in an individual. The new value of the gene is selected randomly from 1 to $SenseNum$, where $SenseNum$ is the maximum number of possible senses for the corresponding word.

- Two parent selection methods are considered: the *roulette wheel* (or *fitness proportionate*) and the *tournament selection* methods. The Sigma scaling function can be used with the roulette wheel selection method to make the GA less susceptible to premature convergence. It is described as follows:

$$ExpVal(i,t) = \begin{cases} 1 + \frac{Fitness(i) - \overline{Fitness}(t)}{2 \cdot \sigma(t)} \\ \text{if } \sigma(t) \neq 0 \\ \\ 1.0 \\ \text{otherwise} \end{cases} \tag{3}$$

$$RWS(i,t) = \frac{ExpVal(i,t)}{\sum_{j=1}^{n} ExpVal(j,t)} \tag{4}$$

where $ExpVal(i,t)$ is the expected value of individual i at iteration t, $RWS(i,t)$ is the probability of individual i to be selected by the roulette wheel at iteration t, $\overline{Fitness}(t)$ is the mean fitness of the population at iteration t, $\sigma(t)$ is the standard deviation of the population fitnesses at iteration t, and n is the population size.

- The *elitist* survivor selection method is considered as a combination between a generational and steady state schemes. The best sequence is then retained unaltered at each generation, which is found generally to significantly improve the performance of GAs.

- Two termination conditions are considered: number of generations and number of fitness evaluations.

6 Experiments

In this section, we present results of experiments with GAWSD on an Arabic data corpus used in [38,78]. It contains 1132 text documents collected from Arabic newspapers [1] from August 1998 to September 2004. This corpus was used for Arabic classification tasks. It contains 6 different categories of documents (arts: 233 documents, economics: 233 documents, politics: 280 documents, sports: 231 documents, woman: 121 documents, and information technology: 102 documents). In our experiments, we selected 60 documents (10 documents from each class) from which we collected 5218 words. With support of linguists, the corpus was manually sense-tagged using AWN. The corpus contains about 48528 sense-tagged instances, which gives an average number of senses per word of 9.3. Two groups of annotators were asked to select the sense for the target word they find the most appropriate in each sentence. The selection of a sense for a target word was made from a list of senses given by AWN. The agreement rate for a target word was estimated as the number of sentences which are assigned identical sense to the target word by the two groups of annotators over the total number of sentences containing the target word. The average inter-annotator agreement gave a score of 91%.

We considered words within a text window to limit the context size (e.g. a window size of 2 means that the context of every word contains at most 5 words, including the target word). These data were used to evaluate the performance of GAWSD under different settings and to compare it with a naïve Bayes classifier. All the results were averaged over 100 runs, and the sense proposed by the algorithm was compared to the manually selected sense.

6.1 Performance evaluation criteria

The performance evaluation criteria were based on the number of True positives (TP), True negatives (TN), False positives (FP), and False negatives (FN).

The fitness evaluation criteria were as follows. The best fitness value $maxFitness$ and its occurrence number nb($maxFitness$) were recorded in each run. The mean fitness $\overline{Fitness}$ and its standard deviation $\sigma(Fitness)$ were calculated over 100 runs.

The performance evaluation criteria were as follows.

1. The precision P is the percentage of correct disambiguated senses for the ambiguous word: $P = TP/(TP+FP), TP+FP \neq 0$.

2. The recall R is the number of correct disambiguated senses over the total number of senses to be given: $R = TP/(TP+FN), TP+FN \neq 0$.

3. The fall-out F is the number of incorrect disambiguated senses over the total number of incorrect senses: $F = FP/(FP+TN), FP+TN \neq 0$.

The results are shown as convergence graphs of the algorithms in respective experiments (parameter setting, impact of parent selection methods, impact of fitness evaluation function, performance evaluation, and comparison of the algorithms). Detailed results of performance comparison are also reported in Tables in terms of mean precision $\overline{P}$, mean recall $\overline{R}$, and mean fall-out $\overline{F}$, along with their respective standard deviations $\sigma(P)$, $\sigma(R)$, and $\sigma(F)$ over the corpus.

The next Section presents the results of experiments conducted on GAWSD for parameters' tuning.

[1] ElAhram: http://www.ahram.org.eg/,
ElAkhbar: http://www.akhbarelyom.org.eg/,
and ElGomhoria: http://www.algomhuria.net.eg/

6.2 Selection of the parameters

The choice of the parameters is critical for the performance of GAs. The first set of experiments involves the numerical investigation of the GA parameters and their impact on the performance of GAWSD, namely, Population size $Population_{size}$, Crossover rate $P_{crossover}$, Mutation rate $P_{mutation}$, and Termination condition $T_{condition}$.

In all experiments on parameters tuning, we used the same following settings: a window size of 2, a random initialization of the population, the roulette wheel as a parent selection method, the Lesk measure as a fitness evaluation function, and $T_{condition} = 50$ generations (except when varying the termination condition).

6.2.1 Variation of population size

We studied the effect of population size on the performance of GAWSD. We chose $P_{crossover} = 0.9$, $P_{mutation} = 0.1$, and made the population size varying from 6 to 100. As shown in Figure 1 (a), the number of best fitness is increasing with the size of the population, but its value is relatively constant. $Population_{size} = 50$ is a good compromise between the number of best fitness and mean fitness.

6.2.2 Variation of crossover and mutation rates

The performance of GAs is always sensitive to the genetic operators' rate. In order to study the effect of varying $P_{crossover}$ and $P_{mutation}$ on the performance of GAWSD. We carried out experiments on the test suite while setting $Population_{size} = 50$. The results are presented in Figure 1 (b,c). The average best results were obtained with $P_{crossover} = 0.70$ and $P_{mutation} = 0.15$.

6.2.3 Variation of termination condition

We studied the performance of GAWSD according to the number of fitness evaluations which we made varying from 1000 to 10,000. The other parameters $Population_{size}$, $P_{crossover}$, $P_{mutation}$ were fixed to 50, 0,70, and 0.15, respectively. Figure 1 (d) shows the results obtained. As expected, the number of best fitness is increasing with the number of fitness evaluations. However, the best fitness value is, in average, more or less indifferent to the number of fitness evaluations starting from 4000.

6.3 Effect of the initial population generation and sigma scaling function

The results of this section are intended to show the combined effect of the method used to generate the initial population (denoted Rnd: randon generation; Cve: constructive generation) and the sigma scaling function (denoted Sig) as a smoothing function for the roulette wheel selection method.

The best fitness is given by all the four variants: random or constructive generation of the initial population, with or without control of the selection pressure of parents (sigma scaling). The highest number of best fitness and best mean fitness are given when the initial population is generated randomly without using the sigma scaling function (Rnd). The second best mean fitness is also obtained with the random generation of the population, when the sigma scaling function is applied ($RndSig$).

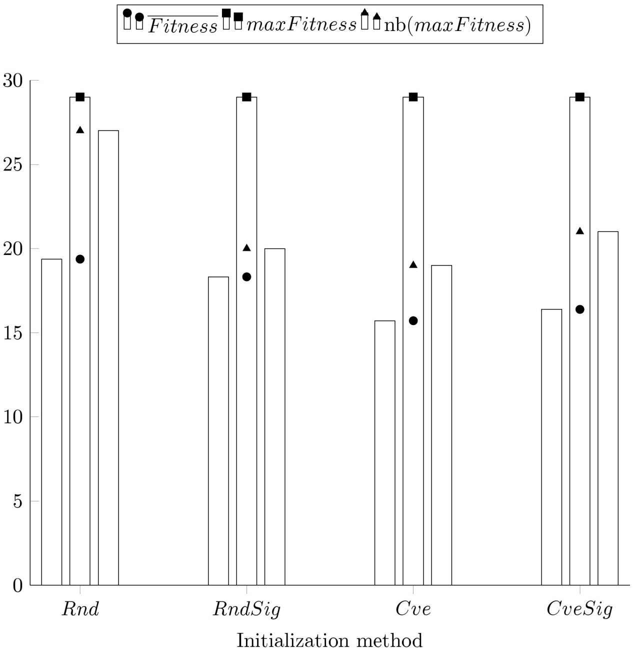
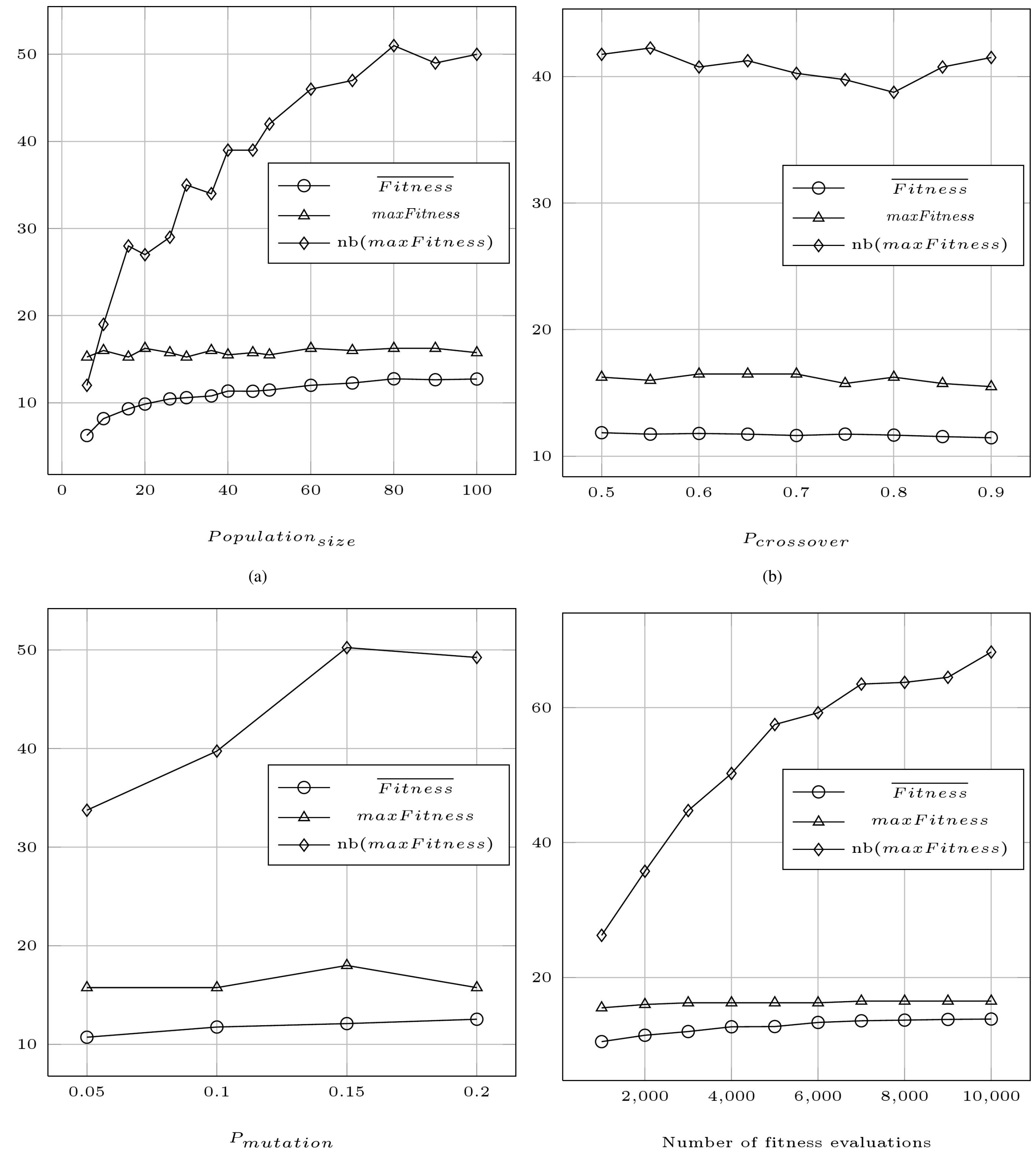


Figure 2: Effect of the initial population generation and sigma scaling function. The abbreviations Rnd, $RndSig$, Cve, and $CveSig$ stand for random generation, random generation + sigma scaling, constructive generation, and constructive generation + sigma scaling, respectively. The best, mean, and number of best fitness values are depicted over 100 runs.

A conclusion can be drawn about the parameter settings. The overall results presented in Figures 1 and 2, substantiate our choice of the following parameters and initialization method for the next experiments: $P_{crossover} = 0.70$, $P_{mutation} = 0.15$, $Population_{size} = 50$, and $T_{condition} \geq 4000$ fitness evaluations. The random generation of the initial population without sigma scaling function is chosen, since it gave substantial improvement with respect to the other schemes.

6.4 Effect of parent selection method

We first investigated the sensitivity of the tournament selection method to variations of tournament size by experimenting with different tournament sizes ($k = 10\% \ldots 40\%$ of the population size $Population_{size}$). Results, presented in Figure 3 (a), show how the best fitness and its mean change with the tournament size. However, the number of best fitness is, in average, constant. The overall best results were obtained with $k = 20$.



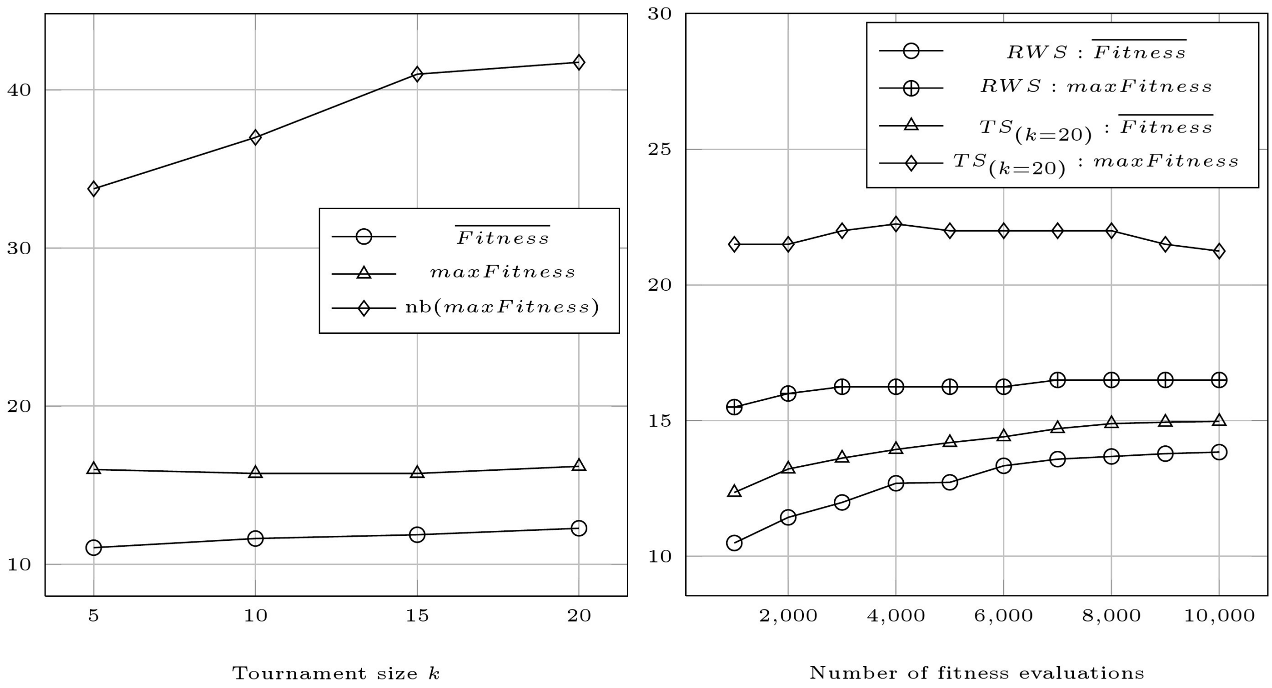
(a)

(b)

(c)

(d)

Figure 1: Selection of the parameters for the algorithm GAWSD. The best, mean, and number of best fitness values are depicted over 100 runs. (a) Variation of the population size $Population_{size}$. (b) Variation of the crossover rate $P_{crossover}$. (c) Variation of the mutation rate $P_{mutation}$. (d) Variation of the termination condition $T_{condition}$.



(a) Variation of the tournament size k.

(b) Parent selection methods.

Figure 3: Comparison of the roulette wheel selection (RWS) and tournament selection (TS). The best, mean, and number of best fitness values are depicted over 100 runs.

We then compared the two parent selection methods: roulette wheel selection (RWS) and tournament selection (TS). The tournament size k was fixed at 20. Figure 3 (b) shows that for the same number of fitness evaluations, the best fitness and its mean obtained with $TS_{k=20}$, were always better than those achieved with RWS.

6.5 Sensitivity to fitness evaluation function

The results of the experiments presented in this section are intended to show the influence of the fitness evaluation function on the performance of GAWSD. We compared four variants of GAWSD based on the parent selection method (RWS or $TS_{k=20}$) and relatedness measure ($Lesk$ or $extendedLesk$): $RWS\&Lesk$, $RWS\&extendedLesk$, $TS_{k=20}\&Lesk$, and $TS_{k=20}\&extendedLesk$.

The graphs of Figure 4 show the results obtained in terms of mean precision $\overline{P}$. The overall best results were obtained with the Lesk measure and the tournament selection method.

6.6 Performance evaluation

To investigate the sensitivity of GAWSD to variations of target word context, we experimented with different text window sizes ($W_{size} = 1 \ldots 5$). The Lesk measure and tournament selection were adopted in GAWSD (GAWSD$_{TS}$). The superiority of the tournament selection method on the roulette wheel selection is shown in Section 6.4.

Two baseline algorithms, Random and FirstSense, were implemented to compare the performance of GAWSD$_{TS}$. Random algorithm selects randomly a sense for each word among its senses given by AWN. FirstSense algorithm selects the first sense that appears in the list of senses for each word.

Results, presented in Figure 5, show how the performance of GAWSD$_{TS}$ changes with the text window size for a given maximum number of fitness evaluations (set to 4000) and how it compares to the baseline algorithms. Convergence graphs of Figure 5 show that GAWSD$_{TS}$ performs better than the baseline algorithms in terms of mean recall and mean precision. The performance improvement of the algorithm is more substantial with the increase in text window size. This behavior was expected, since adding new words to the context of a target word, results in reducing the set of potential word senses, and hence the size of the search space.

6.7 Comparison with other methods

Existing methods related to Arabic WSD, as presented in Section 3, include SALAAM [25], naïve Bayes classifiers [6], and coordination-based semantic similarity [27]. However, their results cannot be compared directly to our algorithms' results, since those methods are related to different WSD tasks and their results were generated on dif-

Table 1: Performance comparison of the algorithms GAWSD$_{TS}$, naïve Bayes (NB), Random, and FirstSense.

Algorithm	$\overline{P}$	$\sigma(P)$	$\overline{R}$	$\sigma(R)$	$\overline{F}$	$\sigma(F)$
GAWSD$_{TS}$	**0.79**	**0.08**	0.63	0.29	**0.20**	**0.12**
NB	0.66	0.21	**0.68**	**0.24**	0.32	0.31
Random	0.38	0.35	0.31	0.42	0.59	0.40
FirstSense	0.54	0.22	0.48	0.30	0.42	0.37

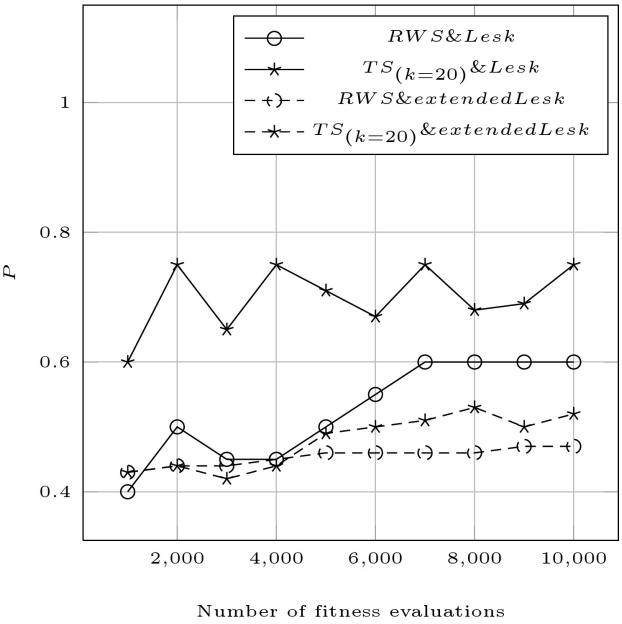


Figure 4: Mean precision of the algorithm GAWSD as a function of the number of fitness evaluations based on the parent selection method (RWS or $TS_{k=20}$) and relatedness measure ($Lesk$ or $extendedLesk$).

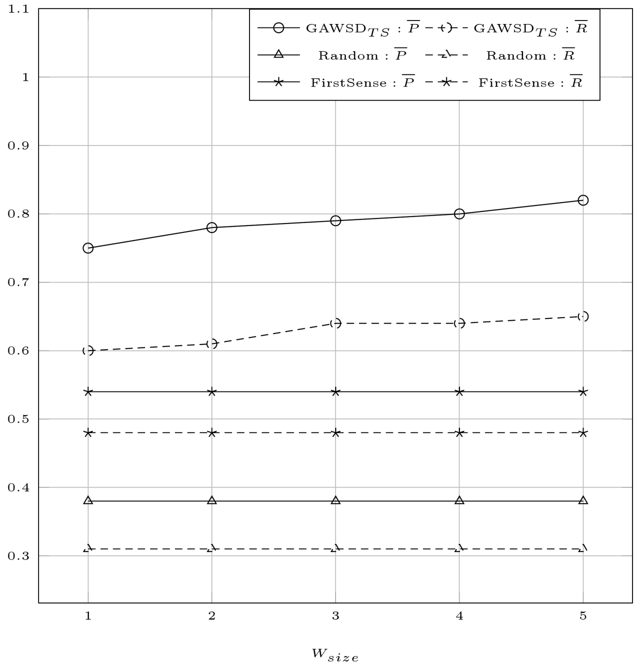


Figure 5: Mean precision and mean recall of the algorithms GAWSD$_{TS}$, Random, and FirstSense as functions of the window size.

ferent corpora. For that reason, we implemented a naïve Bayes classifier to compare its results against those given by GAWSD$_{TS}$. Table 1 presents the average results obtained by GAWSD$_{TS}$, naïve Bayes (NB), Random, and FirstSense algorithms on our corpus. The results show that the best mean precision is given by GAWSD$_{TS}$ ($\overline{P} = 0.79$). Moreover, the standard deviation of the precision ($\sigma(P) = 0.08$) demonstrates its relative robustness. Although, the best mean recall is obtained by the naïve Bayes classifier ($\overline{R} = 0.68$, $\sigma(R) = 0.24$), the mean recall of GAWSD$_{TS}$ is not significantly different ($\overline{R} = 0.63$, $\sigma(R) = 0.29$). This means that GAWSD$_{TS}$ is not only able to find more relevant word senses than the naïve Bayes classifier, but can return most relevant ones as well.

7 Discussion

We evaluated the performance of different variants of GAWSD on a set of 5218 words extracted from an Arabic corpus. The obtained results show that GAWSD$_{TS}$ is the best performing algorithm.

The results of GAWSD$_{TS}$ consistently exhibited superior performance compared with a naïve Bayes classifier and baseline algorithms. Much better precision and recall were obtained by other methods for more specific WSD tasks in Arabic, such as finding correct sense of query translation terms [6], and disambiguating polysemous and homograph Arabic nouns [27].

The results obtained with GAWSD$_{TS}$ in Arabic corroborate those obtained in previous studies on GAs for WSD in Spanish [35] and English [22,84], even though they are not comparable. They confirm that GAs represent a promising approach to WSD and particularly suitable for WSD in Arabic. Indeed, GWSD [84] evaluated on SemCor (English corpora) [58] achieved a mean recall of 71.96%, and GAMBL [22] evaluated on Senseval-3 English all-words task achieved a mean accuracy of 65.2%.

The GA component of the algorithm GAWSD is language-independent. Therefore, GAWSD can be easily adapted to solve WSD in other languages by using specific preprocessing and dictionary, given that a text in the target language can be transformed into a bag of words.

8 Conclusion and future work

This study shows that only few research work has been conducted on WSD problem in Arabic. Indeed, many successful methods have not been investigated yet for Arabic language, comparatively to other natural languages.

We have proposed an evolutionary approach to the WSD problem and applied it to an Arabic corpus. Several variants of the algorithm GAWSD were formulated and examined experimentally on a large set of words extracted from an Arabic corpus. They were assessed on the task of identifying AWN word senses, attaining 79% precision and 63% recall for GAWSD$_{TS}$. Our experiments showed that GAWSD$_{TS}$ outperformed a naïve Bayes classifier in terms of mean precision, which means that GAWSD$_{TS}$ found more relevant word senses than the naïve Bayes classifier. However, their performances in terms of mean recall were comparable, with a small advantage to the naïve Bayes classifier.

Finally, this study opens other directions for future work. The tuning of the parameters remains a major issue to optimize the performance of the proposed algorithms. The results obtained can be improved by implementing a self-adaptive GAWSD that adjusts its parameters during runtime. Furthermore, examining other methods for tuning selection pressure, and thereby exploration/exploitation tradeoff of the algorithms, can have a positive impact on the performance of GAWSD. Another important avenue of research is a thorough study of memetic algorithms for WSD, since they have outperformed GAs on several hard optimization problems.

Acknowledgements

The author gratefully thanks the anonymous referees for their constructive comments on the manuscript.

References

[1] S. Abney, and M. Light (1999). Hiding a semantic class hierarchy in a markov model. In *Proceedings of the ACL Workshop on Unsupervised Learning in Natural Language Processing*, pages 1–8.

[2] E. Agirre, and D. Martinez (2001). Learning class-to-class selectional preferences. In *Proceedings of the 2001 workshop on Computational Natural Language Learning - Volume 7*, ConLL '01, pages 3:1–3:8, Stroudsburg, PA, USA.

[3] E. Agirre, and A. Soroa (2009). Personalizing pagerank for word sense disambiguation. In *Proceedings of the 12th Conference of the European Chapter of the Association for Computational Linguistics*, EACL '09, pages 33–41, Stroudsburg, PA, USA.

[4] H. Al-Serhan, R. Al-Shalabi, and G. Kannan (2003). New approach for extracting arabic roots. In *Proceedings of the 2003 Arab conference on Information Technology*, ACIT'2003, pages 42–59.

[5] R. Al-Shalabi, G. Kanaan, M. Yaseen, B. Al-Sarayreh, and N. Al-Naji (2009). Arabic query expansion using interactive word sense disambiguation. In *Proceedings of the Second International Conference on Arabic Language Resources and Tools*, Cairo, Egypt.

[6] F. Amed, and A. Nürnberger (2008). Arabic/english word translation disambiguation using parallel corpora and matching schemes. In *Proceedings of the 12th Annual Conference of the European Association for Machine Translation*, EAMT 2008, pages 6–11.

[7] L. Araujo (2008). Evolutionary parsing for a probabilistic context free grammar. In *Revised Papers from the Second International Conference on Rough Sets and Current Trends in Computing*, RSCTC '00, pages 590–597, London, UK, Springer-Verlag.

[8] L. Araujo (2002). Part-of-speech tagging with evolutionary algorithms. In *Proceedings of the Third International Conference on Computational Linguistics and Intelligent Text Processing*, CICLing '02, pages 230–239, London, UK, Springer-Verlag.

[9] L. Araujo (2007). How evolutionary algorithms are applied to statistical natural language processing. *Artif. Intell. Rev.*, 28:275–303.

[10] M. Attia (2008). *Handling Arabic morphological and syntactic ambiguities within the LFG framework with a view to machine translation.* PhD thesis, University of Manchester, UK.

[11] S. Banerjee, and T. Pedersen (2003). Extended gloss overlaps as a measure of semantic relatedness. In *Proceedings of the 18th International Joint Conference on Artificial Intelligence*, pages 805–810, San Francisco, CA, USA.

[12] M. Bin-Muqbil (2006). *Phonetic and Phonological Aspects of Arabic Emphatics and Gutturals.* PhD thesis, University of Wisconsin-Madison, WI,USA.

[13] W. Black, S. Elkateb, H. Rodriguez, M. Alkhalifa, P. Vossen, A. Pease, and C. Fellbaum (2006). Introducing the arabic wordnet project. In Fellbaum Sojka, Choi and Vossen eds, editors, *Proceedings of the Third International WordNet Conference*, pages 295–300. Masaryk University, Brno.

[14] S. Bordag (2006). Word sense induction: Triplet-based clustering and automatic evaluation. In *Proceedings of the 11th Conference of the European Chapter of the ACL*, pages 137–144.

[15] J. Brownlee (2011). *Clever Algorithms: Nature-Inspired Programming Recipes*. LuLu.

[16] T. Buckwalter (2004). Buckwalter Arabic Morphological Analyzer (BAMA) version 2.0. Linguistic Data Consortium (LDC), University of Pennsylvania, Philadelphia, PA, USA.

[17] A. Chalabi (1998). Sakhr: Arabic-english computer-aided translation system. In *Proceedings of the Third Conference of the Association for Machine Translation in the Americas on Machine Translation and the Information Soup*, AMTA '98, pages 518–521, London, UK, Springer-Verlag.

[18] M. Ciaramita, and M. Johnson (2000). Explaining away ambiguity: learning verb selectional preference with bayesian networks. In *Proceedings of the 18th Conference on Computational Linguistics - Volume 1*, COLING '00, pages 187–193, Stroudsburg, PA, USA.

[19] S. Clark, and D. Weir (2002). Class-based probability estimation using a semantic hierarchy. *Comput. Linguist.*, 28:187–206.

[20] M. Davis, and T. Dunning (1996). Query translation using evolutionary programming for multilingual information retrieval ii. In *Evolutionary Programming*, pages 103–112.

[21] K. De Jong (2006). *Evolutionary computation - a unified approach*. MIT Press.

[22] B. Decadt, V. Hoste, W. Daelemans, and A. den Bosch. GAMBL, genetic algorithm optimization of Memory-Based WSD. In Rada Mihalcea and Phil Edmonds, editors, *Proceedings of the Third International Workshop on the Evaluation of Systems for the Semantic Analysis of Text (Senseval-3)*, pages 108–112.

[23] M. Diab (2009). Second generation tools (AMIRA 2.0): Fast and robust tokenization, pos tagging, and base phrase chunking. In Khalid Choukri and Bente Maegaard, editors, *Proceedings of the Second International Conference on Arabic Language Resources and Tools*, pages 285–288, Cairo, Egypt.

[24] M. Diab (2003). *Word Sense Disambiguation within a Multilingual Framework*. PhD thesis, University of Maryland College Park, USA.

[25] M. Diab (2004). An unsupervised approach for bootstrapping arabic sense tagging. In *Proceedings of the Workshop on Computational Approaches to Arabic Script-based Languages*, Semitic'04, pages 43–50, Stroudsburg, PA, USA.

[26] M. Diab, M. Alkhalifa, S. Elkateb, C. Fellbaum, A. Mansouri, and M. Palmer (2007). Semeval 2007 task 18: Arabic semantic labeling. In *Proceedings of the 4th International Workshop on Semantic Evaluations*, SemEval '07, pages 93–98, Stroudsburg, PA, USA.

[27] K. Elghamry (2006). Sense and homograph disambiguation in arabic using coordination-based semantic similarity. In *Proceedings of AUC-OXFORD Conference on Language and Linguistics*.

[28] G. Escudero, L. Màrquez, and G. Rigau (2000). Boosting applied to word sense disambiguation. In *Proceedings of the 11th European Conference on Machine Learning*, ECML '00, pages 129–141, London, UK, Springer-Verlag.

[29] G. Escudero, L. Màrquez, G. Rigau, and J. Salgado (2000). On the portability and tuning of supervised word sense disambiguation systems. In *Proceedings of the Joint SIGDAT Conference on Empirical Methods in Natural Language Processing an Very Large Corpora*, pages 172–180.

[30] A. Farghaly, and K. Shaalan (2009). Arabic natural language processing: Challenges and solutions. *ACM Trans, Asian Lang. Inform. Process.*, 8:14:1–14:22.

[31] C. Fellbaum (1998). *WordNet: An Electronic Lexical Database*. Bradford Books.

[32] L. Fogel, A. Owens, and M. Walsh (1966). *Artificial Intelligence through Simulated Evolution*. John Wiley, New York, USA.

[33] W. Gale, K. Church, and D. Yarowsky (2004). Using bilingual materials to develop word sense disambiguation methods. In *Proceedings of the International Conference on Theoretical and Methodological Issues in Machine Translation*, pages 101–112.

[34] M. Galley, and K. McKeown (2003). Improving word sense disambiguation in lexical chaining. In *Proceedings of the 18th international Joint Conference on Artificial Intelligence*, IJCAI'03, pages 1486–1488, San Francisco, CA, USA.

[35] A. Gelbukh, G. Sidorov, and S. Han (2003). Evolutionary approach to natural language word sense disambiguation through global coherence optimization. *WSEAS Transactions on Communications*, 1:11–19.

[36] T. Gharib, M. Habib, and Z. Fayed (2009). Arabic text classification using support vector machines. *International Journal of Computers and Their Applications*, 16(4):192–199.

[37] A. Gliozzo, B. Magnini, and C. Strapparava (2004). Unsupervised domain relevance estimation for word sense disambiguation. In *Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing*, EMNLP, pages 380–387.

[38] N. Habash, and O. Rambow (2005). Arabic tokenization, part-of-speech tagging and morphological disambiguation in one fell swoop. In *Proceedings of the 43rd Annual Meeting on ACL*, ACL '05, pages 573–580, Stroudsburg, PA, USA.

[39] S. Harabagiu, G. Miller, and D. Moldovan (1999). WordNet 2 – a morphologically and semantically enhanced resource. In *Proceedings of the ACL SIGLEX Workshop: Standardizing Lexical Resources*, pages 1–8.

[40] G. Hirst, and D. St Onge (1998). Lexical chains as representation of context for the detection and correction malapropisms. In *WordNet: An Electronic Lexical Database, C. Fellbaum, Ed.*, pages 305–332. MIT Press, Cambridge, MA.

[41] J. Holland (1975). *Adaptation in natural and artificial systems*. University of Michigan press, Ann Arbor, Cambridge, MA, USA.

[42] J. Jiang, and D. Conrath (1997). Semantic similarity based on corpus statistics and lexical taxonomy. http://www.citebase.org/abstract?id=oai:arXiv.org:cmp-lg/9709008

[43] B. Keller, and R. Lutz (1997). Evolving stochastic context-free grammars from examples using a minimum description length principle. In *Workshop on Automata Induction, Grammatical Inference and Language Acquisition (ICML097)*.

[44] S. Khoja (2003). Stemmer. http://zeus.cs.pacificu.edu/shereen/research.htm#stemming

[45] J. Koza (1992). *Genetic programming*. MIT Press, Cambridge, MA, USA.

[46] M. Lapata, and F. Keller (2007). An information retrieval approach to sense ranking. In *Human Language Technologies 2007: Proceedings of the Conference of the North American Chapter of the ACL*, pages 348–355, Rochester, New York.

[47] C. Leacock, G. Miller, and M. Chodorow (1998). Using corpus statistics and wordnet relations for sense identification. *Comput. Linguist.*, 24:147–165.

[48] Y. Lee, and H. Ng (2002). An empirical evaluation of knowledge sources and learning algorithms for word sense disambiguation. In *Proceedings of the ACL-02 Conference on Empirical Methods in Natural Language Processing - Volume 10*, EMNLP '02, pages 41–48, Stroudsburg, PA, USA.

[49] E. Lefever, V. Hoste, and M. De Cock (2011). Parasense or how to use parallel corpora for word sense disambiguation. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies: Short Papers - Volume 2*, HLT '11, pages 317–322, Stroudsburg, PA, USA.

[50] M. Lesk (1986). Automatic sense disambiguation using machine readable dictionaries: how to tell a pine cone from an ice cream cone. In *Proceedings of the 5th Annual International Conference on Systems Documentation*, SIGDOC '86, pages 24–26, New York, NY, USA, ACM.

[51] H. Li, and N. Abe (1998). Generalizing case frames using a thesaurus and the MDL principle. *Comput. Linguist.*, 24:217–244.

[52] D. Lin (1998). An information-theoretic definition of similarity. In *Proceedings of the Fifteenth International Conference on Machine Learning*, ICML '98, pages 296–304, San Francisco, CA, USA.

[53] J. Mallery (1988). *Thinking about foreign policy: Finding an appropriate role for artificial intelligence computers*. PhD thesis, MIT Political Science Department, Cambridge, MA, USA.

[54] D. McCarthy, R. Koeling, J. Weeds, and J. Carroll (2004). Finding predominant word senses in untagged text. In *Proceedings of the 42nd Annual Meeting on ACL*, ACL '04, pages 280–287, Stroudsburg, PA, USA.

[55] Z. Michalewicz (1994). *Genetic algorithms + data structures = evolution programs (2nd, extended ed.)*. Springer-Verlag New York, Inc., New York, NY, USA.

[56] R. Mihalcea (2004). Co-training and self-training for word sense disambiguation. In Hwee, editor, *HLT-NAACL 2004 Workshop: Eighth Conference on Computational Natural Language Learning*, pages 33–40.

[57] R. Mihalcea, P. Tarau, and E. Figa (2004). Pagerank on semantic networks, with application to word sense disambiguation. In *Proceedings of the 20th International Conference on Computational Linguistics*, COLING '04, pages 1126–1132, Stroudsburg, PA, USA.

[58] G. Miller, C. Leacock, R. Tengi, and R. Bunker (1993). A semantic concordance. In *Proceedings of the Workshop on Human Language Technology*, HLT '93, pages 303–308, Stroudsburg, PA, USA.

[59] S. Mohammad, and G. Hirst (2006). Determining word sense dominance using a thesaurus. In *Proceedings of the 11th Conference on European Chapter of the ACL*, EACL, pages 121–128.

[60] R. Mooney (1996). Comparative experiments on disambiguating word senses: an illustration of the role of bias in machine learning. In *Proceedings of the 1996 Conference on Empirical Methods in Natural Language Processing*, pages 82–91.

[61] R. Navigli, and M. Lapata (2010). An experimental study of graph connectivity for unsupervised word sense disambiguation. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 32(4):678–692.

[62] R. Navigli (2009). Word sense disambiguation: A survey. *ACM Comput. Surv.*, 41:10:1–10:69.

[63] R. Navigli, and P. Velardi (2005). Structural semantic interconnections: a knowledge-based approach to word sense disambiguation. *IEEE Trans. Pattern Anal. Mach. Intell.*, 27:1075–1086.

[64] D. Palmer (2002). Text pre-processing. In Nitin Indurkhya and Fred J. Damerau, editors, *Handbook of Natural Language Processing, Second Edition*. CRC Press, Taylor and Francis Group, Boca Raton, FL.

[65] P. Pantel, and D. Lin (2002). Discovering word senses from text. In *Proceedings of the Eighth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '02, pages 613–619, New York, NY, USA.

[66] T. Pedersen (1998). *Learning Probabilistic Models of Word Sense Disambiguation*. PhD thesis, Southern Methodist University, Dallas, TX, USA.

[67] T. Pedersen, and R. Bruce (1997). Distinguishing word senses in untagged text. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP-97)*, pages 197–207, Providence, RI.

[68] R. Rada, H. Mili, E. Bicknell, and M. Blettner (1989). Development and application of a metric on semantic nets. *IEEE Transactions on Systems, Man, and Cybernetics*, 19(1):17–30.

[69] I. Rechenberg (1994). *Evolutionsstrategie'94*, volume 1 of *Werkstatt Bionik und Evolutionstechnik*. Friedrich Frommann Verlag, Stuttgart.

[70] P. Resnik (1993). *Selection and information: a class-based approach to lexical relationships*. PhD thesis, University of Pennsylvania, Philadelphia, PA, USA.

[71] P. Resnik (1995). Using information content to evaluate semantic similarity in a taxonomy. In *Proceedings of the 14th International Joint Conference on Artificial Intelligence - Volume 1*, pages 448–453, San Francisco, CA, USA.

[72] P. Resnik (1997). Selectional preference and sense disambiguation. In *Proceedings of the ACL SIGLEX Workshop on Tagging Text with Lexical Semantics: Why, What, and How?*, pages 52–57.

[73] R. Roth, O. Rambow, N. Habash, M. Diab, and C. Rudin (2008). Arabic morphological tagging, diacritization, and lemmatization using lexeme models and feature ranking. In *Proceedings of the 46th Annual Meeting of the Association for Computational Linguistics on Human Language Technologies: Short Papers*, HLT-Short'08, pages 117–120, Stroudsburg, PA, USA.

[74] M. Sawalha, and E. Atwell (2008). Comparative evaluation of arabic language morphological analysers and stemmers. In *Proceedings of 22nd International Conference on Computational Linguistics*, COLING (Posters), pages 107–110, Manchester, UK.

[75] H. Schütze (1998). Automatic word sense discrimination. *Comput. Linguist.*, 24:97–123.

[76] H-P. Schwefel (1965). *Cybernetic Evolution as Strategy for Experimental Research in Fluid Mechanics (in German)*. PhD thesis, Hermann Föttinger-Institute for Fluid Mechanics, Technical University of Berlin.

[77] M. Sussna (1993). Word sense disambiguation for free-text indexing using a massive semantic network. In *Proceedings of the Second International Conference on Information and Knowledge Management*, CIKM '93, pages 67–74, New York, NY, USA, ACM.

[78] M. Syiam, Z. Fayed, and M. Habib (2006). An intelligent system for arabic text categorization. *International Journal of Intelligent Computing and Information Sciences*, 6(1):1–19.

[79] G. Tsatsaronis, M. Vazirgiannis, and I. Androutsopoulos (2007). Word sense disambiguation with spreading activation networks generated from thesauri. In *Proceedings of the 20th International Joint Conference on Artifical Intelligence*, IJCAI'07, pages 1725–1730, San Francisco, CA, USA.

[80] S. M. van Dongen (2000). *Graph Clustering by Flow Simulation*. PhD thesis, University of Utrecht, The Netherlands.

[81] Z. Wu, and M. Palmer (1994). Verb semantics and lexical selection. In *32nd. Annual Meeting of the Association for Computational Linguistics*, pages 133 – 138, New Mexico State University, Las Cruces, New Mexico.

[82] D. Yarowsky (1994). Decision lists for lexical ambiguity resolution: application to accent restoration in spanish and french. In *Proceedings of the 32nd Annual Meeting on ACL*, ACL '94, pages 88–95, Stroudsburg, PA, USA.

[83] D. Yarowsky (1995). Unsupervised word sense disambiguation rivaling supervised methods. In *Proceedings of the 33rd Annual Meeting on ACL*, ACL '95, pages 189–196, Stroudsburg, PA, USA.

[84] C. Zhang, Y. Zhou, and T. Martin (2008). Genetic word sense disambiguation algorithm. In *Proceedings of the 2008 Second International Symposium on*

Intelligent Information Technology Application - Volume 01, IITA '08, pages 123–127, Washington, DC, USA.

[85] Z. Zhong, and H.T. Ng (2010). It makes sense: a wide-coverage word sense disambiguation system for free text. In *Proceedings of the ACL 2010 System Demonstrations*, ACLDemos '10, pages 78–83, Stroudsburg, PA, USA.

Smart-Home Energy Management in the Context of Occupants' Activity

Domen Zupančič
Robotina d.o.o., OIC-Hrpelje 38, SI-6240 Kozina, Slovenia
Jožef Stefan International Postgraduate School, Jamova 39, SI-1000 Ljubljana, Slovenia
E-mail: domen.zupancic@ijs.si

Božidara Cvetković
Department of Intelligent Systems, Jožef Stefan Institute, Ljubljana, Slovenia
Jožef Stefan International Postgraduate School, Jamova 39, SI-1000 Ljubljana, Slovenia
E-mail: boza.cvetkovic@ijs.si

Keywords: HVAC, energy saving, occupants' comfort, occupants' activity monitoring

Received: June 15, 2014

Energy consumption and occupants' comfort are key factors when evaluating smart-home environments. This paper focuses on occupants' comfort, which is affected by environmental factors (such as temperature, humidity, radiation of elements, and air movement), and occupant-related factors (such as occupants' level of activity, clothing insulation). To satisfy a thermal comfort objective, energy is needed for heating and cooling, which affects energy consumption. This paper presents a proof-of-concept analysis of smart-home control based on occupants' activity level in terms of human energy expenditure, and a trade-off analysis of the energy consumption versus thermal comfort when the activity level serves as an input into an intelligent home energy management system.

Povzetek: V članku je predstavljen inteligentni nadzor pametnega doma, ki temelji na kompromisu med porabo električne energije in udobjem uporabnika.

1 Introduction

Research has focused on regulation of the smart-home environment from various perspectives, including economic, ecological, and assistive. It is common to take into account the occupants' satisfaction and comfort, both depending on environment- and occupant-related factors. The regulation of occupants comfort is a complex and multivariate problem and, to simplify the problem, researchers have assigned static values to occupant-related factors [1, 2, 3].

The complexity of the problem can be expressed as follows. The occupants' comfort must be evaluated according to the knowledge about occupants' activity level and clothing rate and according to the environmental state, such as temperature and humidity. The heating process and its delay due to a room's thermal inertia must be taken into account. The problem is multivariate; the indoor temperature is affected by various heating bodies, such as heat produced by the occupant, the sun through the window, and mechanical heaters [4]. Occupant activity levels and clothing rates can also change much faster than the environmental state. Finally, the effect of the regulation affects comfort slowly and it takes time for an occupant to actually feel this change.

We implemented an agent-based control system that is able to process large data-sets from various external and data sources (weather station, environmental sensors and virtual sensors) and use the extracted knowledge for heating, ventilation and air-conditioning (HVAC) system control to provide occupants with a high level of comfort. We have also deployed a virtual sensing agent for monitoring the activity of occupants in order to provide additional information crucial for regulation of the occupants' thermal comfort. The activity of the occupant is estimated in terms of the human energy expenditure (EE), which is expressed in metabolic equivalent of task (MET), where 1 MET is the energy expended at rest.

The paper presents a multi-agent architecture with virtual sensing agents dedicated to monitoring the real-time state of the occupant (activity level, clothing rate, presence) and shows how the different personal lifestyles influence the HVAC energy consumption and comfort experience.

The rest of the paper is structured as follows. Section 2 presents the background on the topics of comfort experience, human energy expenditure and multi-agent control systems. The system architecture in terms of multi-agent system is presented in section 3. The experimental setup is described in section 4 and respective results are provided in section 5. Section 6 concludes the paper with conclusions and discussion.

2 Background

2.1 Thermal Comfort Experience

Thermal comfort experience is the notion of a person's thermal sensation in a conditioned environment. The predicted mean vote (PMV) index, derived by Fanger et al. [5], expresses the thermal sensation on a seven-point scale ranging from -3 to +3, where negative values denote cold sensation and positive values denote warm sensation. The value 0 denotes neutral sensation, which is the target value for indoor air conditioning. The more distant the PMV is from 0, the more cold (if negative) or hot (if positive) the sensation.

PMV is calculated with the following parameters: clothing insulation (clo_{rate} [clo]), human energy expenditure (EE_{rate} [MET]), air temperature (T_{in} [°C]), relative air velocity (v_{ar} [m/s]), relative humidity (RH [%]), and mean radiant temperature (T_{mr} [$°C$]). The units that are important for this research are defined as follows: 1 clo=0.155 m^2K/W and 1 MET =58.2 W/m^2. In contrast to environmental factors, occupant-related factors are harder to perceive and include into a control system. In [6] we demonstrated the effect that indoor temperature on PMV and decided to regulate PMV maintaining the T_{in}.

The predicted percentage dissatisfied (PPD) measure is used for long term comfort evaluation. PPD predicts the percentage of people who are likely to be dissatisfied with the current thermal state of environment. It depends on PMV and transforms the value of comfort in the range of 5% - 100%. The lowest PPD value equals 5%, which is similar to when PMV equals 0 and is interpreted as follows: at least 5% of people are never satisfied with the thermal state of the environment. PPD and PMV indices formulation is internationally standardised in ISO 7730 [7].

PMV index regulation has been researched to some extent. Calvino et al. [1] developed fuzzy controller for PMV regulation. Ciglar et al. [2] and Liang et al. [3] used the model predictive controller for PMV regulation. Experiments were done in a simulated environment and they all assumed the clothing and activity of a person as a static, predefined value.

2.2 Estimation of Energy Expenditure

The cost of physical activity, namely energy expenditure, is usually expressed in metabolic equivalents of task (MET), where 1 MET is defined as the energy expended at rest. MET values range from 0.9 (sleeping) to over 20 in extreme exertion. Table 1 shows activity levels and their corresponding MET values.

There are a range of methods for reliably estimating energy expenditure (EE). EE can be directly measured using approaches such as direct or indirect calorimetry, or doubly labelled water [8]. These methods are expensive and cumbersome for free-living applications. Accessible commercial devices for estimating EE come in the form of one- [9] [10] or multi-sensor wrist-or armbands [11] that can be used in everyday life. They are based on the concept of high correlation between movement of inertial sensors and activity level, which are in some cases learned using machine-learning algorithms. Most of the methods based on machine learning techniques estimate energy expenditure using wearable sensors and seek linear or nonlinear relations between the energy expenditure and the accelerometer outputs. The most basic methods use one accelerometer and one linear regression model. These approaches can be improved by utilising additional regression models. The method by Crouter et al. [12] uses data from one accelerometer attached to the hip to classify the type of activity (sitting, ambulatory activity or lifestyle activity). According to the recognised type of activity, an appropriate estimation regression model is used, except for sitting, for which a static value of 1 MET is assigned. The drawback of this method is that it can underestimate sedentary activities such as sitting, since such activities are usually accompanied by additional movements (such as office work). Previous research has accelerometers attached to the fixed position on the person's body to bypass the orientation problem of the accelerometer. Our research on estimating EE by using a smartphone in a person's pocket, regardless of orientation, has been shown to product results that are similar or even better [13] than the commercial device SenseWear [11], which is currently claimed to be the most accurate device for free-living situations [14].

Intesity	MET values
Low	EE < 3
Moderate	3 > EE < 6
Vigorous	EE > 6

Table 1: Corresponding MET values for different activity intensity levels.

2.3 Multi-agent control system

Modern buildings contain a range of systems, such as HVAC, domestic hot water system, lighting, safety, etc. Intelligent operation of such systems requires the collection and processing of large sets of heterogeneous data about sensor states, actuator actions, and occupant actions. Distributing tasks among devices, such as smartphones, can provide several types of benefits, such as distributed task solving as well as adding or removing systems and devices during system run-time. The multi-agent system (MAS) approach makes system decentralisation possible. The comparison between the traditional and agent approach was done by Wagner et al. [15], who argued that the agent approach results in a transparent software structure and dynamic and adaptive application software. Klein et al. [16] implemented MAS for coordination of occupant behaviour for building energy and comfort management. Dounis et al. [17] conducted a review on conventional and advanced control systems and implemented MAS for comfort and energy management in buildings, and stated that

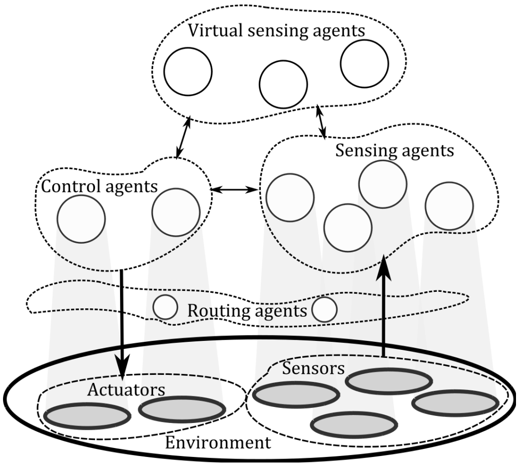


Figure 1: Architecture layers: The bottom layer is the environment with sensors and actuators. The middle layer are the routing agents. The top layer is the sensing, virtual sensing and control agents. The direction of communication is represented with arrows.

a controller has the characteristics of an intelligent agent. Moroşan et al. [18] compared the computational efficiency of centralised and distributed architectures and concluded that the distributed architecture is less computationally demanding than the centralizsed architecture, but achieves the same effect. Our previous research included the development of MAS control architecture, where the operation of each agent was formulated [19].

3 System Architecture

We implemented the smart-home control system as a multi-agent control system. Several agents are developed to operate in groups and each agent can communicate with any other agent. A simplified system architecture is presented in Figure 1. The bottom layer in Figure 1 is the environment including sensors and actuators. Routing agents are above the environment and are used for communication between physical elements of the environment and software agents. The top layer consists of the sensing agents, the control agents and the virtual sensing agents. A group of sensing agents are used to perceive simple environmental states, such as temperature, humidity, acceleration, etc., which can be measured directly, without complex pre-processing. A group of virtual sensing agents is used to perceive complex environmental states that utilise the data obtained from sensing agents or other virtual sensing agents. Control agents are used to affect the environment by changing control parameters and set-point values based on control algorithms, which are denoted as control behaviours. The roles of individual entity behaviour are described in the following subsections.

3.1 Environment

The environment in our research is implemented as a simulator of a building with integrated HVAC equipment. It collects weather data, contains an occupant and generates environmental states. These are represented as state variables and are perceived through sensors. Actions on the environment are performed through actuators. One-way arrows in Figure 1 represent the direction of precipitation and action.

For simulation purposes, we assume the time is represented as a discrete value k, $k \in [0, N-1]$, where N is the number of minutes in the simulation. In each simulation time-step, the environment accepts the vector of set-point values, represented as $\vec{r}(k) = [r_1(k), r_2(k), ...r_K(k)]$ and outputs the vector of state variable values, represented as $\vec{s}(k) = [s_1(k), s_2(k), ...s_J(k)]$ for J environment variables.

3.2 Sensing Agents

Sensor agents are software entities that are used to serve current and historic state variables, obtained through sensors. Sensing agents also include meta-data regarding the physical sensor they represent, such as location of sensor, sensor type, accuracy, drift, unit output, conversion factor, etc. Sensor agents return value of state variable $s(k)$, either on request or based on contract about periodical report between engaged agents using agent communication language (ACL) messaging. The process of contract assignment and cancellation is described in [19].

3.3 Virtual Sensing Agents

Virtual sensor agents are used to perceive and serve current and historic state variables, obtained through sensing agents and/or virtual sensing agents. These agents are used to estimate complex environmental states $cs(k)$ in time k, which cannot be obtained utilising only physical sensors. Examples of environmental states are PMV, PPD, and EE, where the additional processing has to be performed. The processing is based on models that define the relation between environmental states. Detailed functionalities of virtual sensing agents utilised for experiments are presented in the following subsections.

3.3.1 Occupancy Detection Agents

Occupancy detection agents detect the occupancy state of a building based on the data retrieved from sensing agents. For the purposes of this paper, the occupancy state was simulated-agent obtained data about occupancy from a predefined data-set. The approach described by Lu et al. [20] uses a combination of passive infrared motion (PIR) sensors installed in rooms and magnetic reed switches on doors to detect occupancy and sleeping. This approach is inexpensive, unobtrusive, simple to install, and can be used for occupancy detection to extract the occupancy state.

3.3.2 Activity-Monitoring Agents

The activity-monitoring agents communicate with sensing agents to determine current state of the occupant in terms of human energy expenditure (EE). Acceleration data received from sensing agents is collected into 10-second windows, each of which overlaps with the previous one by one half of its length. Each overlapping window is processed into a set of features forming a feature vector that is fed into the regression model to estimate the EE. To build an EE regression model, we have performed two subtasks: (i) machine-learning algorithm selection, and (ii) feature selection, to optimise the performance of the estimation and reduce computational load of the agent.

Selection of appropriate machine-learning regression algorithm was performed with 10-fold cross-validation on the data of 10 people, where one person represents a fold. Data for one person contains regular everyday activities such as rest, cleaning, cooking, and office work, and sporting activities such as walking, running, and stationary cycling. Reference EE expenditure was measured in a controlled environment using indirect calorimetry equipment Cosmed K4b2 [21]. The Table 2 presents the comparison results of regression machine-learning algorithms, Support Vector Regression (SVR), Linear Regression (LR), M5Rules, M5P and REPTree as implemented in Weka machine-learning suite [22]. The results are expressed in mean absolute error (MAE) calculated with Equation 1. We have chosen the SVR algorithm to be deployed in the activity-monitoring agent, since it performs with the lowest estimation error. The best-performing model was compared against the commercial device SenseWear, proving that the smartphone model is comparable to a dedicated device. Note that activity-monitoring agents can use any tri-axial inertial sensor for the EE estimation.

$$MAE = \frac{1}{n}\sum_{1}^{n} |MET_{true} - MET_{predicted}| \,. \qquad (1)$$

Table 2: Results of regression machine-learning algorithms expressed in MAE and comparison against commercial system SenseWear.

Algorithm	MAE
SVR	0.83
LR	0.88
MLP	1.04
M5Rules	1.05
M5P	1.04
REPTree	1.01
SenseWear	0.86

The feature selection procedure was performed using the ranking algorithm RelifF, which ranks the features and assigns each a weight. We have selected only positively weighed features for the final feature set. We began with 67 features computed from acceleration signal and ended with 43 features. The remaining features are partially adopted from Tapia [23] (25 features) and partially developed by us (18 features). Adopted features are: mean of absolute signal value, cumulative sum over absolute signal value, quartiles, variance, inter quartile range, correlation and mean crossing rate. Features developed by us are: signal peak count, cumulative sum over peak absolute value, cumulative sum over signal absolute value, cumulative sum over signal absolute value after band-pass filtering, cumulative square sum over signal absolute value after band-pass filtering, cumulative sum of square components, square of cumulative sum of components after band-pass filtering, velocity, kinetic energy, vector length, integration of area under vector length curve. The highest-ranked features are quartiles (four features) and peak count (one feature).

3.3.3 Clothing Detection Agent

The clothing detection agent communicates with the activity-monitoring agent and sensing agents to predict the type of clothing a user is currently wearing. The output is expressed in unit clo, where one clo is the amount of insulation that allows a person at rest to maintain thermal equilibrium in an environment at 21°C.

The prediction is based on simple heuristics that utilise information about the current season, current weather, time of the day, and estimated EE in the indicated order. Examples of the rules can be observed in Rules 1 and 2.

Rule 1:

```
if $season$ is $winter$ and $weather$ is $sunny$
then if $time > 11\ PM$
then if $EE_{rate} > 2\ MET$
then $clo_{rate} = 1$
else $clo_{rate} = 2$
```

Rule 2:

```
if $season$ is $winter$ and $weather$ is $sunny$
then if $time > 7\ AM$ and $time < 11\ PM$
then if $EE_{rate} > 3\ MET$
then $clo_{rate} = 0.5$
else $clo_{rate} = 1$
```

Rule 1 predicts the clo_{rate} value according to the amount of activity in the evening, where a higher clo_{rate} value indicates higher thermal insulation due to clothes or blankets. Rule 2 predicts the clo_{rate} value according to the estimated EE during the day. If the occupant's EE_{rate} is higher than 3 MET, this indicates exercise.

3.3.4 Comfort Estimation Agent

A comfort estimation agent is used to perceive the state of comfort, expressed as PMV according to ISO 7730, Annex D [7]. For such purpose, it obtains values T_{in} and RH from sensing agents and clo_{rate} and EE_{rate} from the clothing detection agent and the activity monitoring agent, respectively, for the current time k. The T_{mr} is assumed equal T_{in} and v_{ar} is assumed fixed as 0.15 m/s.

3.4 Control agent

The control agent includes an algorithm for defining indoor temperature set-point values T_s in order to regulate PMV when the building is occupied. If the building is not occupied, the T_s is fixed at 5°C and 40°C for the heater and the chiller, respectively, in order to prevent freezing or overheating of HVAC components. There are two versions of control agents: the heater control agent and the chiller control agent. The desired range PMV_{range} includes acceptable PMV values and is specified with a value of PMV_{ref} so that $PMV_{range} \in [-PMV_{ref}, +PMV_{ref}]$. The control algorithm is designed in order to increment/decrement the set-point value of T_s in a way that brings the PMV value on the borders of PMV_{range}. The $-PMV_{ref}$ value is a target value for the PMV for the heater control agent and $+PMV_{ref}$ for the chiller control agent. The increment function for set-point temperature T_s for both versions is defined as:

$$T_s(k+1) = T_s(k) + T_{inc}(k), \tag{2}$$

where the set-point temperature in the next time step $T_s(k+1)$ is computed according to the previous set-point temperature $T_s(k+1)$, incremented by a value $T_{inc}(k)$. T_{inc} is computed according to the Equation 3. The PMV_{diff} definition is presented in Equation 4.

$$\begin{aligned} T_{inc}(k) = A \cdot PMV_{diff}(k)^3 + \\ + B \cdot \frac{PMV_{diff}(k)}{D \cdot T_{diff}(k)^2 + 1} + \\ + C \cdot PMV_{diff}(k) \end{aligned} \tag{3}$$

$$PMV_{diff} = \begin{cases} -PMV_{ref} - PMV, & \text{if heater} \\ +PMV_{ref} - PMV, & \text{if chiller} \end{cases} \tag{4}$$

The PMV_{diff} expresses the distance between the PMV and the value PMV_{ref} in case of chiller and the PMV_{diff} expresses the distance between the PMV and the value $-PMV_{ref}$ in the case of the heater, where the T_{diff} definition is presented in Equation 5.

$$T_{diff}(k) = T_s(k) - T_{in}(k) \tag{5}$$

Equation 3 represents the regulation algorithm, which is proportional to the difference between the current PMV and PMV_{ref} and tends to achieve the equality of mentioned values, which is the purpose of our algorithm. Furthermore, Equation 3 is inverse-proportional to the difference between the current T_{in} and current T_s with the purpose of reducing the fluctuation of the T_s value in order to reduce instability of regulation system. The constant values $A = 0.01$, $B = 3$, $C = 0.1$ and $D = 1$ were obtained iteratively with several simulation runs in order to achieve the desired control effect.

4 Experimental Setup

Experiment consists of two individual procedures: the EE model creation and evaluation in isolations already presented in Section 3.3.2, and evaluation of the model on two-day data of the occupants.

Data used in the first experiments was collected in a laboratory environment, where the persons performed predefined scenarios (rest, cooking, cleaning, walking, running, cycling). They carried a smartphone in their trouser pockets, from which we collected the raw acceleration data. For reference energy expenditure measurements, the Cosmed K4b2 indirect calorimeter was used. Both acceleration data and Cosmed data were synchronised to produce a training and testing data-set.

Since we could not collect the free-living reference dataset, due to the cumbersome nature of Cosmed, we have created synthetic two-day data with a one-minute sampling rate from the synchronised recordings (smartphone and Cosmed) for five people used in this experiment. This data-set was further enriched with weather data (including outdoor temperature, humidity, wind speed and solar radiation). The weather data was collected from the Slovenian Environment Agency (ARSO) portal [24] and represents two sunny days in February 2014, city Rateče.

The data represents one working day and one non-working day (e.g., a weekend) for each person. The characteristics of their lifestyles can be observed in Table 3. The goal was to produce data for people with different lifestyles. They are summarised as follows. Persons B, C and D have regular eight-hour jobs, where Person A works at night and Person E works from home. Person A does regular exercise such as walking and vigorous running on a treadmill. Person B does regular exercise on weekends. Person C is engaged in very intensive home chores over both days. Person D is an athlete and frequently exercises vigorously. Person E does not exercise and leaves home only for half an hour. Table 3 shows the percentage of time the occupant was at home, the absolute minutes of performed activities with low, moderate, and vigorous intensity. The new data-set were included into a simulation environment.

Experimental setup included the control system, developed using JADE [25]; the simulation model, developed using EnergyPlus software [26]; and the simulation environment, developed using BCVTB [27] software. Machine-learning models for energy expenditure were implemented using Weka [22].

We have instantiated sensor agents for acceleration, indoor temperature, mean radiant temperature, relative humidity and outdoor temperature. We then instantiated the activity-monitoring agent, comfort estimation agent, clothing detection agent, and occupancy detection agent. Fi-

		Intensity (hours)		
		Low	Moderate	Vigourous
Person	At home (%)	< 3	$3 \geq\leq 6$	> 6
A	54	20.8	2.1	2.8
B	71	32.2	1.2	0.7
C	56	22.9	4.1	0
D	67	27.7	2.9	1.9
E	99	45.5	1.9	0

Table 3: The characteristics of the occupants for the produced days. The percentage of time present at home and amount of activities in hours according to the intensity (MET).

nally, two instances of control agents - one heater and one chiller agent - were created.

A simulation model of a building containing one room was taken into account for the experiment; this model was obtained from the EnergyPlus software project examples and represents a thermal dynamic model of the room. The room has an integrated packaged terminal heat pump with chiller, heater and supplementary heater rated at 8500W, 8000 W and 3000 W respectively as a HVAC system, together with the temperature regulation module. The heating power produced by a person P_{pHeat} is computed according to Equation 6.

$$P_{pHeat} = EE_{rate} \cdot 58.2W/m^2 \cdot A_{body} \cdot phr \qquad (6)$$

A_{body} is the area of a human body, phr is the person heat rate. We took the $phr = 0.8$, which indicates that 80% of energy, consumed by a person is transformed to heat and $A_{body} = 1.8m^2$, as computed for an average person of weight 70 Kg and height 1.73 m [5] using the Dubios equation [28]. Table 4 presents some values of heating power, produced by a person at various EE_{rate}.

Table 4: Heating power produced by a human body at various EE_{rate} values

EE_{rate} [MET]	P_{pHeat} [W]
1	83.81
2	167.62
3	251.42
4	335.23
5	419.04
6	502.85
7	586.66
8	670.46

A simulation time-step was set to one minute. Each simulation time-step a simulation environment - BCVTB - (i) accepts temperature set-point variables from control agents, (ii) computes new environmental states based on EnergyPlus model and (iii) passes them to sensing agents. One routing agent was instantiated for variable mapping between the simulation environment - BCVTB and JADE agents.

5 Results

5.1 Energy Expenditure Estimation

The activity-monitoring agent processed the collected acceleration data for respective occupant returning the estimation of current energy expenditure. Results in terms of MAE for each occupant are presented in Figure 2. We can observe that the agent performs with low error in case of low- and moderate-intensity activities (rest, home chores, walking) and slightly underestimates more vigorous activities (cycling and running). This shortcoming can be solved by employing activity recognition as a part of activity-monitoring agents and utilising multiple regression models according to the specific activity.

5.2 Maintaining Comfort and Energy Consumption

We performed 41 simulations on the synthetic data-set explained in section 4. A simulation starts with $PMV_{ref} = 0.10$ and each further simulation has PMV_{ref} incremented for 0.01 until the value PMV_{ref}=0.40 (41st simulation run). Note that 80% of energy consumed by a person was transformed to heat as defined in simulation model.

Simulation of part of a day (200 minutes) is presented on Figure 3. It shows the controller's performance when the intensity of the activity changes significantly. Between the 2200th and 2220th minutes, the occupants' intensity of activity decreases from 2.2 MET to 1.3 MET. PMV decreases immediately and the controller starts increasing the T_{in} as seen around the 2220th minute. It increases the T_{in} until the PMV reaches value 0 in the 2230th minute, where a small overshoot can be seen (0.5°C difference between the values of T_{in} in 2230th and 2250th). Afterwards, the controller handles small changes of PMV due to small changes in the intensity of activity. Between the 2260th and 2320th minutes, the intensity of activity changes rapidly for approximately 1 MET. We can observe that the controller could not handle such changes on time, so in that period the PMV fluctuates between +1 and -1. In the next period (until the 2340th minute), the controller increases the temperature and PMV is stabilised. Finally, in the period between the 2340th and 2380th minutes, we can observe more efficient handling of PMV since the PMV is not fluc-

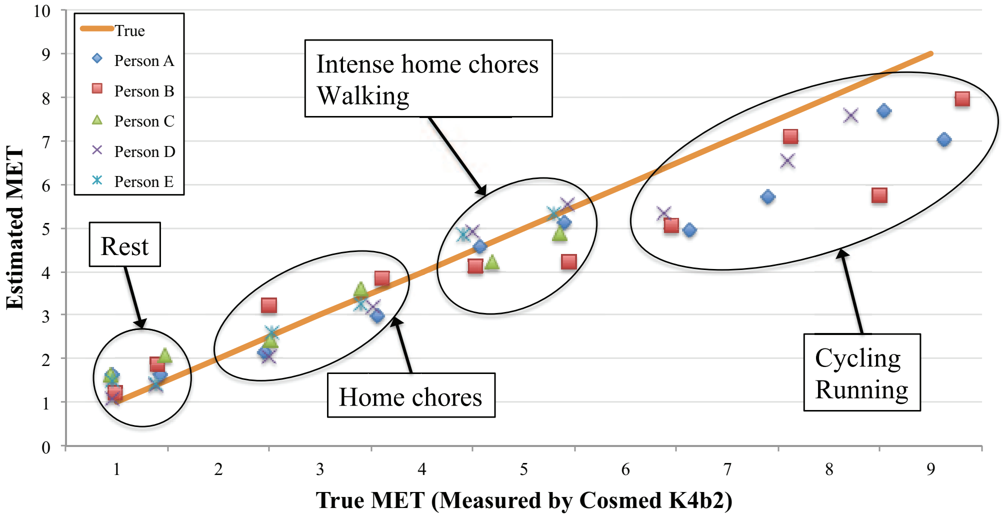


Figure 2: True MET vs. Estimated MET for each occupant. The utilised regression model performs with low MAE on low and moderate activities and slightly underestimates the vigorous activities.

tuating greatly. It changes between +0.2 and +0.7, which denotes a slightly warm sensation.

Each simulation run returns the energy consumed for HVAC and the average value of PPD during occupancy. The average value $\overline{PPD}$ is computed using Equation 7. The individual time-step is denoted with k, the number of time-steps with N, and $occ(k)$ represents the occupancy state (0 or 1) as returned by the occupancy detection agent.

$$\overline{PPD} = \sum_{k=1}^{N} PPD(k)\ \forall k,\ occ(k) = 1 \qquad (7)$$

The obtained comfort/consumption plane is presented in Figure 4. Simulation results for one occupant are presented with the same marker. The right-most marker of each series is a simulation result when PMV_{ref} is 0.1. Higher values of PMV_{ref} shift result markers to the left, giving the left-most marker when PMV_{ref} is 0.4. Lower values of $\overline{PPD}$ and energy denotes better comfort experience and lower energy consumption. We can observe that Person C obtains the lowest values for both objectives, thus producing the best result. This is related to the person's low home presence rate (56%), as shown in Table 3. A low "At Home" rate implies lower energy consumption and 0 hours of vigorous actives implies that the control system did not need to handle severe PMV changes. The worst results are obtained for persons A and D, who have a lower "At Home" rate, but a higher EE_{rate}. A higher EE_{rate} results in higher $\overline{PPD}$, in this case above 20%, which indicates a low overall comfort rate. Persons B and E return average results. Their "At Home" rate is higher, which indicates high energy consumption for low PMV_{ref} values compared to other persons. High-intensity activities of Person B, compared to Person E, result in a worse overall comfort experience.

Figure 5 compares different parameter configuration for Person E, again simulated for 41 PMV_{ref} values. The estimated EE_rate and the clo_rate is a result of the estimated EE_{rate} and estimated clo_{rate}. The fixed EE_rate and the clo_rate denote 1.2 MET and 1 clo fixed during the entire simulation run, as seen in related work [2]. The fixed clo_rate is a result of the estimated EE_{rate} and the fixed clo_{rate}. The fixed EE_rate is a result of the fixed EE_{rate} and estimated clo_{rate}.

We can observe that using fixed values for EE_{rate} and clo_{rate} makes the regulation underestimate the $\overline{PPD}$ and consumes much more energy compared to the estimated values of EE_{rate} and clo_{rate}.

The fixed clo_{rate} implies higher energy consumption, but a similar range of $\overline{PPD}$. In such a case, the energy consumed for heating when the occupant is asleep is higher due to lower clothing insulation (1 clo) compared to the estimated value clo_{rate}, which is 2 clo when sleeping. A similar effect occurs when a person exercises. If the clo_{rate} is fixed, the energy consumed for cooling when the occupant exercises is higher due to higher clothing insulation (1 clo) compared to estimated value clo_{rate}, which is 0.5 clo when the occupant exercises.

The fixed EE_{rate} also makes the regulation underestimate the $\overline{PPD}$ and consume more energy than the estimated values of EE_{rate} and clo_{rate}. Energy consumption is lower than the fixed values of both EE_{rate} and clo_{rate}. In that case, clothing insulation reduces the energy consumption for heating (when occupant sleeps) and for cooling (when occupant exercises).

6 Conclusion and discussion

This paper presents the multi-agent system for HVAC, which regulates the occupant's comfort according to activity level, clothing rate, and occupant's presence. We argue that our dynamic treatment of occupants' comfort enables better comfort and lower energy consumption than activity-

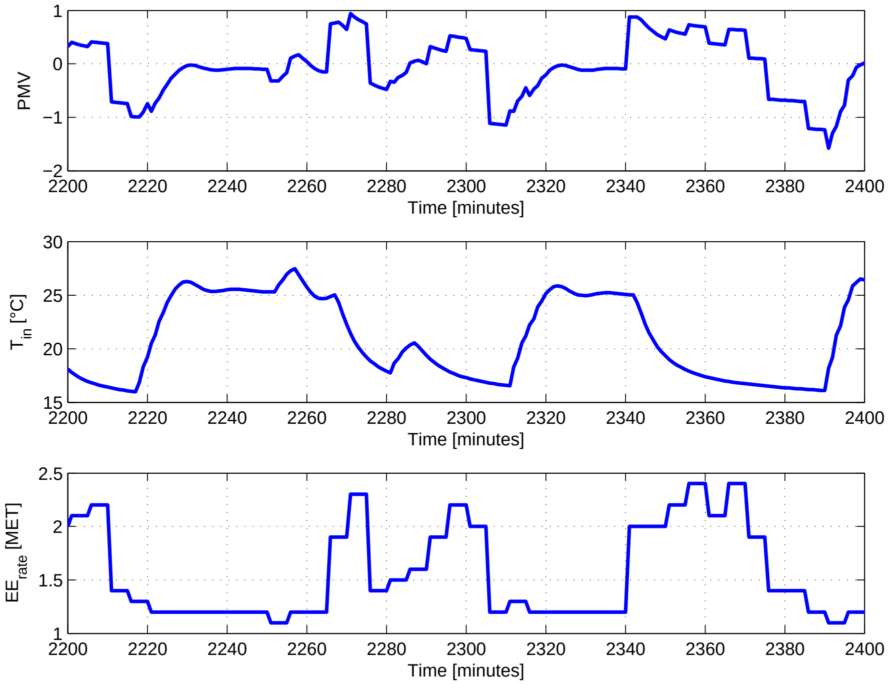


Figure 3: A section of a day, when a person performs low-intensity activities. Top figure: PMV value, Middle figure: indoor temperature, Bottom figure: EE_{rate}

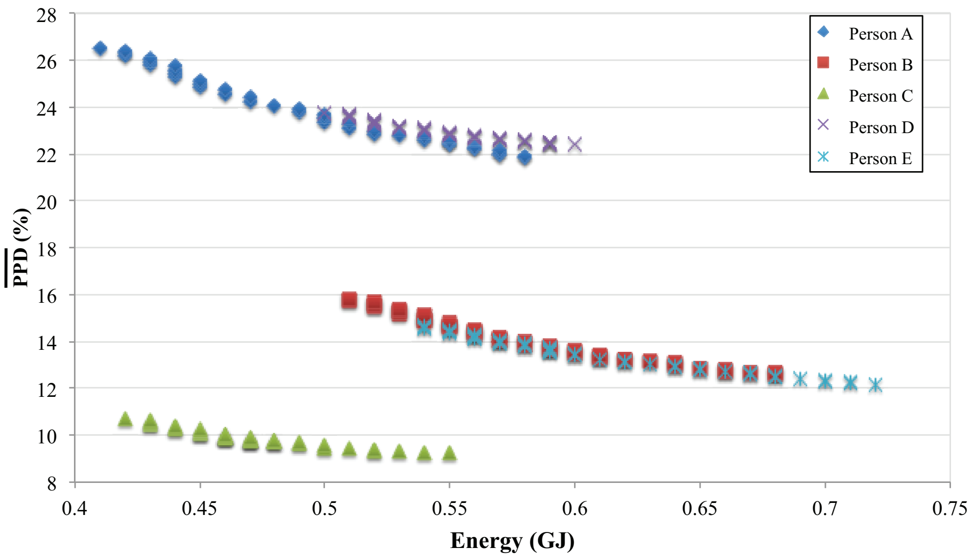


Figure 4: Comfort/consumption plane: the X axis represents the energy in GJ, consumed by HVAC; the Y axis represents the comfort experience - $\overline{PPD}$, expressed in $\%$. Each marker presents a result of a simulation run for different PMV_{ref} and for different person

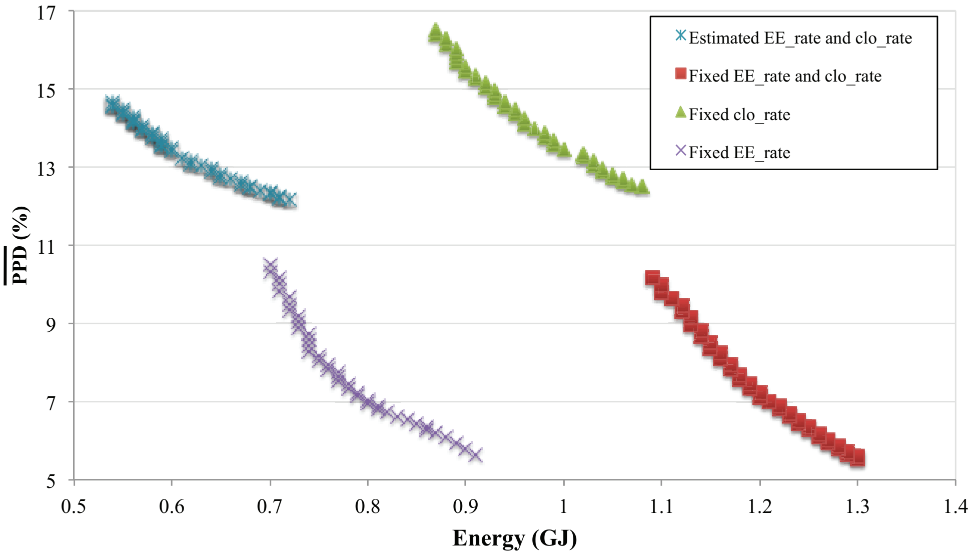


Figure 5: Comfort/consumption plane: The X axis represents the energy in GJ, consumed by HVAC; the Y axis represent the comfort experience - $\overline{PPD}$, expressed in %. Each marker presents a result of a simulation run for different PMV_{ref} and for a different assumption of the occupant's parameters configuration

and clothing-independent treatment of comfort presented in related work.

The multi-agent system consists of three virtual sensing agents. The first is the activity-monitoring agent, which, in contrast to other related research where activity level is assigned a static value, dynamically estimates the human energy expenditure of the occupant, utilising sensor agents such as smartphone accelerometer data. Second, the clothing detection agent utilises the environmental sensors (weather) and activity-motoring agent to predict the value of clothing isolation, which is adopted in related work as a static value. Third, the comfort estimation agent utilises data from the activity-monitoring agent, the environmental sensing agents (weather, indoor temperature, humidity), and clothing detection agent to estimate the occupant's comfort.

The control agent utilises the information from the sensing agents and virtual sensing agents to regulate comfort (to reach the comfort equilibrium) by maintaining the indoor temperature.

We have shown that our multi-agent system can efficiently regulate the comfort for people with certain lifestyles. We have analysed the trade-off between comfort and energy consumption, which is highly affected by heating objects or energy released by an occupant.

Future work will consist of adapting human energy expenditure estimation model to the specific person [29] and predicting the human energy expenditure, since the estimation contributes to delay in temperature regulation. We will implement the presence classification agent that will predict the occupant's time of arrival. Moreover, it is crucial to improve the control algorithm in order to achieve the quicker response needed to eliminate the delay in regulation.

Acknowledgement

Operation part financed by the European Union, European Social Fund. Operation implemented in the framework of the Operational Programme for Human Resources Development for the Period 2007- 2013, Priority axis 1: Promoting entrepreneurship and adaptability, Main type of activity 1.1.: Experts and researchers for competitive enterprises.

References

[1] F. Calvino, M. La Gennusa, M. Morale, G. Rizzo, and G. Scaccianoce. Comparing different control strategies for indoor thermal comfort aimed at the evaluation of the energy cost of quality of building. *Applied Thermal Engineering*, 30(16):2386 – 2395, 2010. Selected Papers from the 12th Conference on Process Integration, Modelling and Optimisation for Energy Saving and Pollution Reduction.

[2] J. Cigler, S. Prívara, Z. Váňa, E. Žáčeková, and L. Ferkl. Optimization of predicted mean vote index within model predictive control framework: Computationally tractable solution. *Energy and Buildings*, 52(0):39 – 49, 2012.

[3] J. Liang and R. Du. Design of intelligent comfort control system with human learning and minimum power control strategies. *Energy Conversion and Management*, 49(4):517 – 528, 2008.

[4] C. E. Garcia, D. M. Prett, and M. Morari. Model predictive control: theory and practice–a survey. *Automatica*, 25(3):335–348, 1989.

[5] P. Fanger et al. Thermal comfort. analysis and applications in environmental engineering. *Thermal comfort. Analysis and applications in environmental engineering.*, 1970.

[6] D. Zupančič, B. Cvetkovič, and M. Gams. Smart-home energy management system: A trade-off between energy consumption and thermal comfort experience according to occupant's activity. In *Proceedings of the 6th Jožef Stefan International Postgraduate Students Conference*, 2014.

[7] International Standard Organization. ISO 7730. Technical report, International Standards Organization, 2006.

[8] R. Brychta, E. Wohlers, J. Moon, and K. Chen. Energy expenditure: Measurement of human metabolism. *Engineering in Medicine and Biology Magazine, IEEE*, 29(1):42–47, Jan 2010.

[9] Nike+. FuelBand. "`http://www.nike.com/us/en_us/c/nikeplus-fuelband`", 2014.

[10] Fitbit Flex. "`http://www.fitbit.com`", 2014.

[11] Indirect calorimeter. "`http://sensewear.bodymedia.com/`", 2014.

[12] S. E. Crouter, J. R. Churilla, and D. R. Bassett. Estimating energy expenditure using accelerometers. *European Journal of Applied Physiology*, 98(6):601–612, 2006.

[13] B. Cvetković, B. Kaluža, R. Milić, and M. Luštrek. Towards human energy expenditure estimation using smart phone inertial sensors. In *Ambient Intelligence*, volume 8309 of *Lecture Notes in Computer Science*, pages 94–108. Springer International Publishing, 2013.

[14] JM Lee, Y Kim, and GJ Welk. Validity of consumer-based physical activity monitors. *Medicine and science in sports and exercise*, February 2014.

[15] H. Mubarak and P. Gohner. An agent-oriented approach for self-management of industrial automation systems. In *Industrial Informatics (INDIN), 2010 8th IEEE International Conference on*, pages 721–726, July 2010.

[16] L. Klein, J. Kwak, G. Kavulya, F. Jazizadeh, B. Becerik-Gerber, P. Varakantham, and M. Tambe. Coordinating occupant behavior for building energy and comfort management using multi-agent systems. *Automation in Construction*, 22:525–536, 2012.

[17] A. Dounis and C. Caraiscos. Advanced control systems engineering for energy and comfort management in a building environment—a review. *Renewable and Sustainable Energy Reviews*, 13(6):1246–1261, 2009.

[18] P. Moroşan, R. Bourdais, D. Dumur, and J. Buisson. Building temperature regulation using a distributed model predictive control. *Energy and Buildings*, 42(9):1445 – 1452, 2010.

[19] D. Zupančič, M. Luštrek, and M. Gams. Multi-agent architecture for control of heating and cooling in a residential space. *The Computer Journal*, 2014.

[20] J. Lu, T. Sookoor, V. Srinivasan, G. Gao, B. Holben, J. Stankovic, E. Field, and K. Whitehouse. The smart thermostat: Using occupancy sensors to save energy in homes. In *Proceedings of the 8th ACM Conference on Embedded Networked Sensor Systems*, pages 211–224, Zürich, Switzerland, 2-5 November 2010. ACM, New York, USA.

[21] Cosmed k4b2. Indirect calorimeter. "`http://www.cosmed.it/en/products/indirect-calorimetry`", 2014.

[22] M. Hall, E. Frank, G. Holmes, B. Pfahringer, P. Reutemann, and I. H. Witten. The weka data mining software: An update. *SIGKDD Explor. Newsl.*, 11(1):10–18, November 2009.

[23] E. M. Tapia. *Using Machine Learning for Real-time Activity Recognition and Estimation of Energy Expenditure*. PhD thesis, Massachusetts Institute of Technology, 2008.

[24] Slovenian Environment Agency. ARSO. "`http://www.arso.gov.si/en/`", 2014.

[25] F. Bellifemine, F.Bergenti, G. Caire, and A. Poggi. JADE – a java agent development framework. In R.H. Bordini, M. Dastani, and A.E.F. Seghrouchni, editors, *Multi-Agent Programming*. Springer, New York, USA, 2005.

[26] D. B. Crawley and et al. Energyplus: creating a new-generation building energy simulation program. *Energy and Buildings*, 33(4):319–331, 2001.

[27] Michael Wetter. Co-simulation of building energy and control systems with the building controls virtual test bed. *Journal of Building Performance Simulation*, 4(3):185–203, 2011.

[28] D. Du Bois and E.F. Du Bois. A formula to estimate the approximate surface area if height and weight be known. *Nutrition (Burbank, Los Angeles County, Calif.)*, 5(5):303, 1989.

[29] B. Cvetković, B. Kaluža, M. Luštrek, and M. Gams. Adapting activity recognition to a person with multi-classifier adaptive training. *Journal of Ambient Intelligence and Smart Environments*, 2014.

The Unexpected Hanging Paradox from an AI Viewpoint

Matjaž Gams
Jozef Stefan Institute, Jamova 39, 1000 Ljubljana, Slovenia
E-mail: Matjaz.Gams@ijs.si, dis.ijs.si/mezi

Position paper

Keywords: logical paradox, artificial intelligence, progress of human civilization

Received: June 6, 2014

This position paper hypothesizes that humans are becoming smarter, not only when using ICT and AI tools, but on their own, particularly due to the progress of AI knowledge. As is common when demonstrating that one computing mechanism is stronger than another, we chose a well-known task – the unexpected hanging paradox –that humans were previously unable to resolve efficiently, but can now do so thanks to new knowledge. We show that the cause of prior problems was with ambiguous definition, as it was in the case of the liar paradox.

Povzetek: Predstavljena je hipoteza, da ljudje postajamo čedalje pametnejši zaradi spoznanj umetne inteligence, pokazana na paradoksu nepričakovanega obešanja.

1 Introduction

According to the Flynn effect [1], scores on the standard broad-spectrum IQ tests improve by up to three IQ points each decade, and the gains are even higher in some specialized areas. One theory claims that the increase of human intelligence is related to the use of information tools [2], which often progress exponentially over time.[3]

This paper presents a tentative hypothesis that artificial intelligence (AI) influences human intelligence in a positive way; specifically, it increases the ability to solve mental problems. We illustrate the hypothesis in Figure 1. The *y* axis is logarithmic in the scale. Therefore, the linear growth of computer skills on the graph corresponds to the exponential nature of Moore's law.[4] Basic human physical and mental properties, such as speed of movement, coordination or speed of human computing, have remained nearly constant in recent decades, as represented by the horizontal line in Figure 1.

Our first thesis is that, analogous to mechanical machines that enable humans to move faster than on their own, the ability of humans to solve problems increases due to information tools such as computers, mobile devices with advanced software, and AI in particular (the bold top line in the Figure 1). (The overall human ability to solve problems is growing, due to a number of reasons, primarily the growth of ICT capabilities, or advances in computers, mobile devices, and the Web.) Programs such as the Google browser may provide the greatest knowledge source available to humans, thereby representing an extension of our brains.

We go a step further in this paper. Whereas mechanical machines do not increase our physical capabilities, human intelligence generally increases on its own. For example, not only does a person play better chess when using advice from online chess programs, they also perform better when playing against other human opponents. This is due to previous interactions with chess-playing programs. In the AI community [5], it is generally accepted that AI progress is increasing and might even enable human civilization to take a quantitative leap.[6]

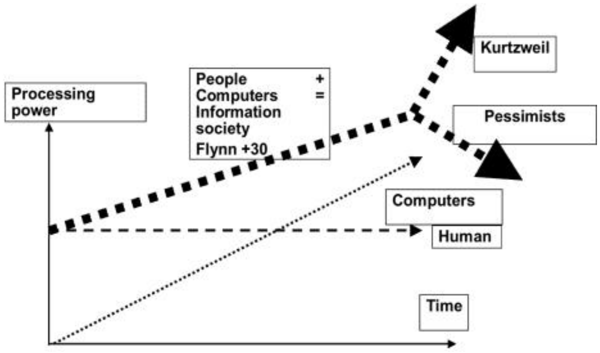


Figure 1: Growth of computer and human computing powers.

Several opposing theories claim that humans actually perform worse on their own, since machines and tools have replaced humans' need to think on their own. We argue that while this effect may be valid for human physical properties such as obesity, it is not the case in mental tasks. Another pessimistic viewpoint suggests that intelligent civilizations decline after reaching a certain development level (see Figure 1), possibly because of overpopulation, self-destruction or depletion of natural resources. This would explain why we have not yet detected alien civilizations, though the Drake's equation [7] indicates that many such civilizations should exist.

In real life, ICT, and AI tools in particular, have already significantly modified the way humans exercise their mental activities. For example, professional chess players intensively use computer chess programs to analyze game strategies and improve their level of play. Furthermore, computers outperform the best humans in nearly all mental games, with some rare exceptions such as Go. Therefore, this online advice helps humans play much better than on their own. Although it is safe to claim that computers have already significantly improved human gaming performance, is the phenomenon valid in other areas?

If we can show that humans can solve logical puzzles that they were not able to solve until recently without computers, that would be a good indication of humans getting smarter on their own. One way to confirm this idea would be to analyze the logical solutions that humans solved in the last decade. Another way would be to provide a new solution to an existing puzzle. One objection might be that just one solution of one puzzle is far too little to show anything. On the other hand, since the author of this paper is a well-educated AI scientist and not a professional logician, it could provide a reasonable indication that the tentative idea might be valid.

To demonstrate the idea, we analyze the unexpected hanging paradox.[8, 9, 10, 11] In addition, we discuss if AI programs would crash from such well-known logical paradoxes or resolve them.

2 The liar paradox

First, however, we quickly investigate the liar paradox (in which a liar says that he is a liar), first published in [12]. According to [13] it was first formulated by the Greek philosopher Eubulides of Miletus: “A man says that he is lying. Is what he says true or false?” This sentence is false when it is true. It supposedly leads to a paradox and causes logical AI machines to crash, such as in the “I, Mudd” episode of the science-fiction television series *Star Trek*.

However, as Prior shows [14], there is no paradox, since the statement is false. It is a simple contradiction of the form “A and not A,” or “It is true and false.” In other words, **if a person always lies by definition, then that person is, by definition, not allowed to say anything that is not a lie**. Therefore, such statements are simply not allowed, which means they are false. In summary, no decent AI computing machine should fail to see the falsity of the liar paradox sentence.

How did the liar paradox cause such attraction? An explanation at hand is that humans fall into a loop of true/untrue derivations without observing that their thinking was already falsified by the declaration of the problem. It seems a valid logical problem, so humans apply logical reasoning. However, the declaration of the logical paradox was illogical at the start rendering logical reasoning meaningless.

In another example, 1 + 1 = 2, and we all accept this as a true sentence without any hesitation. Yet, one liter of water and one liter of sugar do not combine to form two liters of sugar water. Therefore, using common logic/arithmetic in such a task is inappropriate from the start.

The principle and paradox of multiple-knowledge [15] tentatively explain why humans easily resolve such problems. We use multiple knowledge/ways of thinking not only in parallel, but also with several mental processes interacting together during problem-solving. Different processes propose different solutions, and the best one is selected. The basic difference in multiple-knowledge viewpoint compared to the classical ones occurs already at the level of neurons. The classical analogy of a neuron is a simple computing mechanism that produces 0/1 as output. In the multiple viewpoint, each neuron outputs 2^N possible outcomes, which can be demonstrated if *N* outputs from a single neuron are all connected to *N* inputs of another neuron. In summary, the multiple-knowledge principle claims that the human computing mechanism at the level of a neuron is already much more complex than commonly described, and even more so at the level of higher mental processes.

According to the principle of multiple knowledge, humans have no problems computing that one apple and one apple are two apples, and 1 liter of water and 1 liter of sugar is 1.6 liters of liquid and a mass of 2.25 kilograms, since they use multiple thinking. A person who logically encounters the sugar-water merge for the first time may claim that it will result in 2 liters of sugar water. However, after an explanation or experiment, humans comprehend the problem and have no future problems of this kind.

Another AI solution at hand uses contexts. In arithmetic, 1 + 1 = 2. In merging liquids and solid materials, 1 + 1 ≠ 2. In the first case, the context was arithmetic and in the second case, merging liquids and solid materials. The contexts enable an important insight into the unexpected handing paradox.

3 The unexpected hanging paradox

Unlike the liar paradox, the unexpected hanging paradox (also known as the hangman paradox, the unexpected exam paradox, the surprise test paradox, or the prediction paradox) yields no consensus on its precise nature, so a final correct solution has not yet been established.[9] This is a paradox about a person’s expectations about the timing of a future event that they are told will occur at some unexpected time.[16]

The paradox has been described as follows [9]:

A judge tells a condemned prisoner that he will be hanged at noon on one weekday in the following week but that the execution will be a surprise to the prisoner. He will not know the day of the hanging until the executioner knocks on his cell door at noon that day.

Having reflected on his sentence, the prisoner draws the conclusion that he will escape from the hanging. His reasoning is in several parts. He begins by concluding that the "surprise hanging" can't be on Friday, as if he hasn't been hanged by Thursday, there is only one day *left - and so it won't be a surprise if he's hanged on Friday. Since the judge's sentence stipulated that the hanging would be a surprise to him, he concludes it cannot occur on Friday.*

He then reasons that the surprise hanging cannot be on Thursday either, because Friday has already been eliminated and if he hasn't been hanged by Wednesday night, the hanging must occur on Thursday, making a Thursday hanging not a surprise either. By similar reasoning he concludes that the hanging can also not occur on Wednesday, Tuesday or Monday. Joyfully he retires to his cell confident that the hanging will not occur at all.

The next week, the executioner knocks on the prisoner's door at noon on Wednesday — which, despite all the above, was an utter surprise to him. Everything the judge said came true.

Evidently, the prisoner miscalculated, but how? Logically, the reasoning seems correct. While there have been many analyses and interpretations of the unexpected hanging paradox, there is no generally accepted solution. The paradox is interesting to study because it arouses interest in both laymen and scientists. Here, we provide a different analysis based on the viewpoint of cooperating AI agents [16][5], contexts and multiple knowledge.[15]

The prediction of hanging on one out of five possible days is well defined through a real-life empirical fact of a human life being irreversibly terminated. However, the surprise is less clearly defined. If it denotes cognitive surprise, then the prisoner can be sure that the hanging will take place on the current day. No surprise is assured each new day, even on the first day, so hanging under the given conditions is not possible. Such an interpretation makes no sense. To avoid the prisoner being cognitively certain, the following modifications are often proposed [9]:

The prisoner will be hanged next week, and the date (of the hanging) will not be deductible in advance from the assumption that the hanging will occur during the week (A).

The prisoner will be hanged next week and its date will not be deducible in advance using this statement as an axiom (B).

Logicians are able to show that statement (B) is self-contradictory, indicating that in this interpretation, the judge uttered a self-contradicting statement leading to a paradox.

Chow [10] presents a potential explanation through epistemological formulations suggesting that the unexpected hanging paradox is a more intricate version of Moore's paradox [9]:

A suitable analogy can be reached by reducing the length of the week to just one day. Then the judge's sentence becomes: "You will be hanged tomorrow, but you do not know that."

Now we can apply AI methods to analyze the paradox. First, the judge's statement is a one-sided contract (an agreement can always be written in the form of a contract) from an AI agent viewpoint, defining a way of interacting and cooperating. As with any agreement/contract, it also has some mechanisms defining the consequences if one side violates the agreement. Since the judge unilaterally proclaimed the agreement, he can even violate it without any harm to him, whereas the prisoner's violations are punished according to the judge's will and corresponding regulations. For example, if the prisoner harms a warden, the deal is probably off, and the hanging can occur at the first opportunity, regardless of whether it is a surprise. This is an introductory indication that the hanging paradox is from the real world and that it matters, and is not just logical thinking. Even more important, it enables a valid conclusion that **any error in prisoner's actions releases the judge from his promise**.

On the other hand, the judge is, by definition, an honest person and as long as the prisoner abides to the appropriate behavior, the judge will keep his word and presumably postpone the execution if the prisoner predicts the exact day of the hanging. Now, we come to the crucial definition ambiguity. The term *deducible* means that the prediction will be 100 percent guaranteed accurate about a one-time event (that is, hanging), so such **a prediction can be uttered only once a week, not each day anew**. Therefore, the prisoner has exactly one chance of not only predicting, but also **explaining with certainty to the judge**, why the hanging will occur on that particular day. The judge will have to be persuaded; that is, he will have to understand and accept the prisoner's line of reasoning. If not, the deal is off and the judge can choose any day while still keeping his word.

For further understanding of *deducible*, consider a case in which the prisoner is given a life-saving coupon on which he writes the predicted day and stores it in the judge's safe on Monday morning with the explanation attached. Obviously, the prisoner stands no chance if the judge orders handing on Monday. Namely, if the prisoner proposes Monday, he cannot provide a deducible explanation why the handing will happen on Monday. Yes, he will not be surprised in cognitive terms, but both a correct prediction and a deducible explanation are required in order to avoid hanging. The only chance to avoid hanging is to predict Friday and hope that he will not be hanged till Friday. (In this case, the judge could still object that, on Monday for example, the prisoner could not provide a plausible explanation for Friday. Yet, that would not be fair since, on Friday, the prisoner would indeed be sure of the judge coming into contradiction.) Even if the prisoner is allowed to deposit the one and only coupon on any day in the week, there is no major difference in terms of explanation in this paper. Again, if the prisoner is allowed to deposit the coupon each day anew, this formulation makes no sense.

To explain the error in the prisoner's line of reasoning (that is, logical induction), assume that instead of giving his ruling five days in advance, he gave it on Thursday morning, leaving a two-day opportunity. Since the prisoner could use the single pardon (remember: *deducible* for a one-time event means one prediction once) and save himself on Friday, he concludes that Thursday is the only day left and cashes in his only coupon with a 100 percent certain logical explanation on Thursday. However, in this case the judge could carry out the hanging on Friday. Why? Because the prisoner provided the only 100 percent certain prediction in the form of a single life-saving coupon on Thursday, which means that on Friday he could not deliver the coupon. In other words, the prisoner wrongly predicted the hanging day and therefore violated the agreement.

It turns out that the situation on Thursday is similar to the situation on Monday. Even if the judge knocks on the door on Thursday, and the prisoner correctly predicted Thursday, he still could not provide a 100 percent certain explanation why the hanging would occur on Thursday since the judge could come back on Friday as described in the above text; therefore, the judge can proceed on Thursday without violating his proclamation.

What about AI machines? Will they crash or fail as was supposed to be the case with the liar paradox? Similarly to the liar paradox, the principle of multiple knowledge provides a simple solution that AI machines should be able to compute. If both lines of reasoning (from Friday to Monday or from Monday to Friday) are simulated with some tests, the solutions should be obtained. One does not need to understand why one line of reasoning is wrong in order to operationally solve the puzzle. The AI machine can simply evaluate both of them and accept the more plausible one. However, current AI systems are not yet capable of understanding the explanation in this paper since they behave poorly on any task demanding real-life semantics.

4 Discussion

Wikipedia offers the following statement regarding the unexpected hanging paradox [9]:

There has been considerable debate between the logical school, which uses mathematical language, and the epistemological school, which employs concepts such as knowledge, belief and memory, over which formulation is correct.

According to other publications [8], this statement correctly describes the current state of scientific literature and the human mind.

To some degree, solutions similar to the one presented in this paper have already been published.[8–9] However, they have not been generally accepted and, in particular, have not been presented through AI means. Namely, AI enables the following explanation:

The error in the prisoner's line of reasoning occurs when extending his induction from Friday to Thursday, as noted earlier, but the explanation in this paper differs. The correct conclusion about Friday is not:

"Hanging on Friday is not possible" (C),

but :

"**If** not hanged till Friday **and** the single prediction with explanation was not applied for any other day before, **then** hanging on Friday is not possible." (D)

The first condition in (D) is part of common knowledge. The second condition in (D) comes from common sense about one-sided agreements: every breach of the agreement can cause termination of it. An example would be promising a one-sided reward to a person for predicting an outcome of a sporting event and then realizing that the person deposited two predictions.

The two conditions reveal why humans have a much harder time understanding the hanging paradox, compared to the liar paradox. The conditions are related to the concepts of ***time* and *deducibility* and should be applied simultaneously**, whereas only one insight is needed in the liar paradox. In AI, this phenomenon is well known as the **context-sensitive reasoning** (often related to agents), which was first presented in [18] and has been used extensively in recent years. Here, as in real life, under one context the same line of reasoning can lead to a different conclusion compared to the conclusion under another context (remember the sugar water). But one can also treat the conditions in statement (D) as logical conditions, in which case the context can serve for easier understanding. The same applies to the author of this paper: Although he has been familiar with the hanging paradox for decades, the solution at hand emerged only when the insight related to the contexts appeared.

Returning to the motivation for analysis of the unexpected hanging paradox, the example was intended to show that humans have mentally progressed to see the trick in the hanging paradox, similar to how people became too smart to be deceived by the liar paradox.

There are several potential objections. First, one needs no AI or ICT knowledge to see the proposed solution. However, this is the only major change of the author's knowledge from the years before the recent progress of AI knowledge. It is not only that using AI knowledge helped solve the paradox. It also enabled a shift from correct logical thinking under wrong preconditions into multiple, agent- and context-based thinking to avoid the logical trap.

The second objection could be that human civilization has not yet accepted the explanation provided here, and the validity of the hypothesis relies on future acceptance of the explanation. The danger is that humans will ignore or oppose the explanation provided here. If so, consequent disclaimers will have to be published in this journal as well. On the other hand, this is the purpose of scientific position papers.

Third, the proposed solutions to the analyzed logical puzzle might seem to be just one single event and not that the human civilization has improved due to advances in AI, ICT, and cognitive science. However, these and similar paradoxes have stirred human imagination for eons and have not yet been satisfactorily resolved, even by brilliant mathematicians and logicians. In addition, these problems are known globally. Therefore, we must rely on new knowledge when providing the explanation in this paper. Furthermore, in order to show superior computing performance of one mechanism over another, it is necessary to show just one task that a certain mechanism can solve and the other cannot. According to the tentative hypothesis presented here, have we not shown how human mental capabilities have increased in recent decades, since an intelligent individual can understand the solution provided herein but the best knowledge among the smartest individuals could not previously?

This new approach has also been used to solve several other paradoxes, such as the blue-eyes paradox and the Pinocchio paradox. Analyses of these paradoxes are being submitted to other journals.

In summary, the explanation of the hanging paradox and the difficulty for human paradox solvers resembles those of the liar paradox before solving it beyond doubt. It turns out that **both paradoxes are not truly paradoxical**; instead, they describe a logical problem in a way that a human using logical methods cannot resolve the problem. Similar to the untrue assumption that a liar can utter a true statement, the unexpected hanging paradox in the prisoner's line of reasoning exploits **two misconceptions**. The first is that a 100 percent accurate prediction for a single event can be uttered more than once (through a vague definition of "surprise") and the second that a conclusion that is valid at one time is also valid during another time span (moving from Friday to Thursday; that is, not accepting the conditions in statement *C*).

Due to the simplicity of the AI-based explanation in this paper, there is no need to provide additional logical, epistemological, or philosophical mechanisms to explain the failure of the prisoner's line of reasoning. There is nothing wrong with inductive reasoning, as long as preconditions are valid.

The hanging paradox is interesting from various perspectives, such as regarding the question of which methods enable successful analysis and explanation. This paper provides an AI-based explanation for humans, while other explanations, such as an explanation or procedure for AI machines to analyze the unexpected hanging paradox, remain a research challenge.

Acknowledgements

The author wishes to thank several members of the Department of Intelligent Systems, particularly Boštjan Kaluža and Mitja Luštrek, for their valuable remarks. Special thanks are also due to Angelo Montanari, Stephen Muggleton, and Eva Černčic for contributions on this and other logic problems. Also, the anonymous reviewers provided several remarks that helped improve the article.

References

[1] Neisser, U. (1997). Rising scores on intelligence tests. *American Scientist* 85, Sigma Xi, 440–7.

[2] Flynn, J. R. (2009). *What Is Intelligence: Beyond the Flynn Effect*. Cambridge, UK: Cambridge University Press.

[3] Computing laws revisited (2013). *Computer* 46/12.

[4] Moore, G.E. (1965). Cramming more components onto integrated circuits. *Electronics Magazine*, 4.

[5] *Proceedings of the Twenty-Third International Joint Conference on Artificial Intelligence* (IJCAI'13) (2013). Beijing, China.

[6] Kurzweil, R. (2005). *The Singularity is Near*. New York: Viking Books.

[7] Dean, T. (2009). A review of the Drake equation. *Cosmos Magazine*.

[8] Wolfram A. (2014). http://mathworld.wolfram.com/UnexpectedHangingParadox.html

[9] Unexpected hanging paradox, Wikipedia (2014). https://en.wikipedia.org/w/index.php?title=Unexpected_hanging_paradox&oldid=611543144, June 2014

[10] Chow, T.Y. (1998). The surprise examination or unexpected hanging paradox. *American Mathematical Monthly* 105:41–51.

[11] Sober, E. (1998). To give a surprise exam, use game theory. *Synthese* 115:355–73.

[12] O'Connor, D.J. (1948). Pragmatic paradoxes. *Mind* 57: 358–9.

[13] Beall, J.C., Glanzberg, M. (2013). In Edward N. Zalta, E.N. (eds.), *The Stanford Encyclopedia of Philosophy*.

[14] Prior, A.N. (1976). *Papers in Logic and Ethics*. Duckworth.

[15] Gams, M. (2001). *Weak Intelligence: Through the Principle and Paradox of Multiple Knowledge*. New York: Nova Science Publishers, Inc.

[16] Sorensen, R. A. (1988). *Blindspots*. Oxford, UK: Clarendon Press.

[17] Young, H.P. (2007). The possible and the impossible in multi-agent learning. *Artificial Intelligence* 171/7.

[18] Turner, R.M. (1993). Context-sensitive Reasoning for Autonomous Agents and Cooperative Distributed Problem Solving, In *Proceedings of the IJCAI Workshop on Using Knowledge in its Context*, Chambery, France.

A Unified Framework for Detection of Suspicious and Anomalous Beahvior from Spatio-Temporal Traces

Boštjan Kaluža
Department of Intelligent Systems, Jozef Stefan Institute, Jamova cesta 39, Ljubljana, Slovenia
bostjan.kaluza@ijs.si, http://bostjankaluza.net

Thesis Summary

Keywords: agent, agent behavior, suspicious behavior, anomalous beahvior, anomaly detection

Received: June 10, 2014

This paper presents a summary of the doctoral dissertation of the author on the topic of learning patterns of agent behavior from sensor data.

Povzetek: Članek predstavlja povzetek doktorske disertacije avtorja, ki obravnava temo učenja vzorcev obnašanja agenta iz senzorskih podatkov.

1 Introduction

The problem of learning behavior patterns from sensor data arises in many applications including smart environments, video surveillance, network analysis, human-robot interaction, and ambient assisted living. Our focus is on detecting behavior patterns that deviate from regular behaviors and might represent a security risk, health problem, or any other abnormal behavior contingency. In other words, deviant behavior is a data pattern that either does not conform to the expected behavior (anomalous behavior) or matches previously defined unwanted behavior (suspicious behavior). Deviant behavior patterns are also referred to as outliers, exceptions, peculiarities, surprise, misuse, etc. Such patterns occur relatively infrequently; however, when they do occur, their consequences can be quite dramatic, and often negative.

We targets a large class of problems with complex, spatio-temporal, sequential data generated by an entity capable of physical motion in environment, regardless of whether the observed entity is human, software agent, or even robot. In such domains, an agent often has an observable spatio-temporal structure, defined by the physical positions relative to static landmarks and other agents in environment. We suggest that this structure, along with temporal dependencies and patterns of sequentially executed actions, can be exploited to perform deviant behavior detection on traces of agent activities over time.

2 Unified detection framework

We propose a unified framework to analyze agent behavior from prior knowledge and external observations in order to detect deviant behavior patterns. A detailed unified framework flowchart is outlined in Figure 1.

From the behavior analysis perspective, we propose a novel, efficient encoding that we refer to as a spatio-activity matrix. This matrix is able to capture behavior dynamics in a specific time period using spatio-temporal features, whereas its visualization allows visual comparison of different behavior patterns. Next, we provide a feature extraction technique, based on principal component analysis, in order to reduce the dimensionality of the spatio-activity matrix. We then introduce a clear problem definition that helps establish a theoretical framework for detecting anomalous and suspicious behavior from agent traces in order to show how to optimally perform detection. We discuss why detection error is often inevitable and prove the lower error bound, and provide several heuristic approaches that either estimate the distributions required to perform detection or to directly rank the behavior signatures using machine learning approaches. The established theoretical framework is extended to show how to perform detection when the agent is observed over longer periods of time and no significant event is sufficient to reach a decision. We specify conditions that any reasonable detector should satisfy, analyze several detectors, and propose a novel approach, referred to as a F-UPR detector, that generalizes utility-based plan recognition with arbitrary utility functions.

3 Empirical studies

The unified framework is demonstrated in three studies: detection of decreased behavior that indicates disease or deterioration in the health of elderly persons; detection of suspicious passengers in the airport simulation; and verification of persons at an access control point in high-security application.

The first study introduces an approach to monitoring an individual at home by an ambient-intelligence system to detect daily living pattern anomalies. It utilizes the pro-

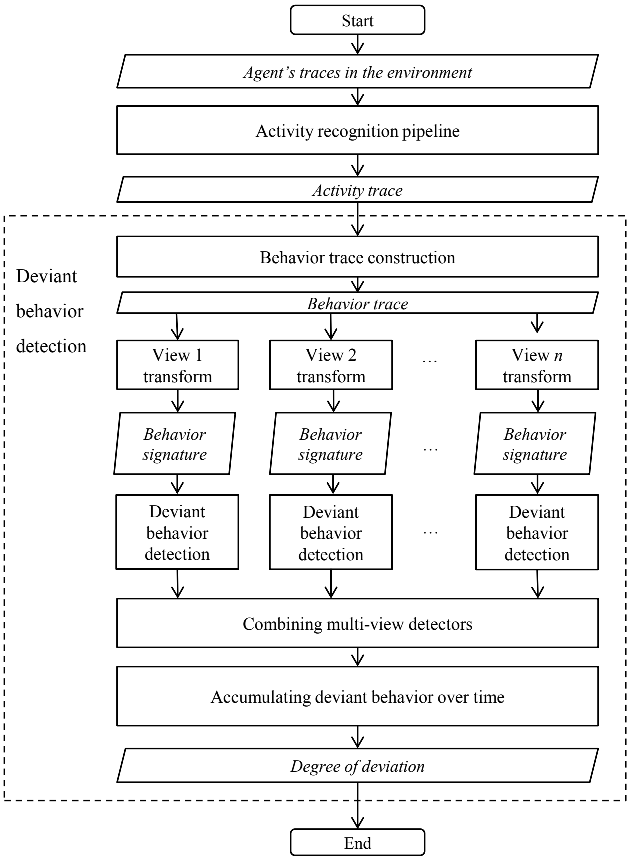

Start
Agent's traces in the environment
Activity recognition pipeline
Activity trace
Deviant behavior detection
Behavior trace construction
Behavior trace
View 1 transform
View 2 transform
View n transform
Behavior signature
Behavior signature
Behavior signature
Deviant behavior detection
Deviant behavior detection
Deviant behavior detection
Combining multi-view detectors
Accumulating deviant behavior over time
Degree of deviation
End

Figure 1: Processing flowchart of the unified framework.

posed unified framework to recognize activities, extract spatio-activity behavior signatures, and apply an outlier-detection method to classify the individual's daily patterns, regardless of the cause of the problem, be it physical or mental. Experiments indicate that the proposed solution successfully discriminates between healthy person behavior patterns and those of a person with health problems.

The second study focuses on two applications in surveillance domain, where the goal is to detect suspicious agents in the environment. In particular, it targets a large class of applications where no single event is sufficient to gauge whether or not agent behavior is suspicious. Instead, we face a sparse set of trigger events that identify interesting parts in behavior trace. The first application considers suspicious passenger detection at an airport, while the second application tackles dangerous driver detection.

The third study concerns entry control, which is an important security measure that prevents undesired persons from entering secure areas. The utilized unified detection framework allows an advanced risk analysis to distinguish between acceptable and unacceptable entries, based on several entry sensors, such as fingerprint readers, and intelligent methods that learn behavior from previous entries. First, it analyzes person behavior from different viewpoints and then performs a joint risk analysis. The obtained results represent an improvement in detecting security attacks.

In summary, we proposed a novel framework for suspicious and anomalous behavior detection, and demonstrated its applicability in three empirical studies.

References

[1] Kaluža, B. Detection of suspicious and anomalous beahavior detection from spatio-temporal agent traces. PhD thesis, Jozef Stefan International Postgraduate School (2013).

[2] Kaluža, B.; Gams, M. Analysis of daily-living dynamics. *Journal of Ambient Intelligence and Smart Enviroments* 4, 403-–413 (2012).

[3] Kaluža, B.; Dovgan, E.; Tušar, T.; Tambe, M.; Gams, M. A probabilistic risk analysis for multimodal entry control. *Expert Systems with Applications* 38, 6696-–6704 (2011).

JOŽEF STEFAN INSTITUTE

Jožef Stefan (1835-1893) was one of the most prominent physicists of the 19th century. Born to Slovene parents, he obtained his Ph.D. at Vienna University, where he was later Director of the Physics Institute, Vice-President of the Vienna Academy of Sciences and a member of several scientific institutions in Europe. Stefan explored many areas in hydrodynamics, optics, acoustics, electricity, magnetism and the kinetic theory of gases. Among other things, he originated the law that the total radiation from a black body is proportional to the 4th power of its absolute temperature, known as the Stefan–Boltzmann law.

The Jožef Stefan Institute (JSI) is the leading independent scientific research institution in Slovenia, covering a broad spectrum of fundamental and applied research in the fields of physics, chemistry and biochemistry, electronics and information science, nuclear science technology, energy research and environmental science.

The Jožef Stefan Institute (JSI) is a research organisation for pure and applied research in the natural sciences and technology. Both are closely interconnected in research departments composed of different task teams. Emphasis in basic research is given to the development and education of young scientists, while applied research and development serve for the transfer of advanced knowledge, contributing to the development of the national economy and society in general.

At present the Institute, with a total of about 900 staff, has 700 researchers, about 250 of whom are postgraduates, around 500 of whom have doctorates (Ph.D.), and around 200 of whom have permanent professorships or temporary teaching assignments at the Universities.

In view of its activities and status, the JSI plays the role of a national institute, complementing the role of the universities and bridging the gap between basic science and applications.

Research at the JSI includes the following major fields: physics; chemistry; electronics, informatics and computer sciences; biochemistry; ecology; reactor technology; applied mathematics. Most of the activities are more or less closely connected to information sciences, in particular computer sciences, artificial intelligence, language and speech technologies, computer-aided design, computer architectures, biocybernetics and robotics, computer automation and control, professional electronics, digital communications and networks, and applied mathematics.

The Institute is located in Ljubljana, the capital of the independent state of **Slove**nia (or S♡nia). The capital today is considered a crossroad between East, West and Mediterranean Europe, offering excellent productive capabilities and solid business opportunities, with strong international connections. Ljubljana is connected to important centers such as Prague, Budapest, Vienna, Zagreb, Milan, Rome, Monaco, Nice, Bern and Munich, all within a radius of 600 km.

From the Jožef Stefan Institute, the Technology park “Ljubljana” has been proposed as part of the national strategy for technological development to foster synergies between research and industry, to promote joint ventures between university bodies, research institutes and innovative industry, to act as an incubator for high-tech initiatives and to accelerate the development cycle of innovative products.

Part of the Institute was reorganized into several high-tech units supported by and connected within the Technology park at the Jožef Stefan Institute, established as the beginning of a regional Technology park "Ljubljana". The project was developed at a particularly historical moment, characterized by the process of state reorganisation, privatisation and private initiative. The national Technology Park is a shareholding company hosting an independent venture-capital institution.

The promoters and operational entities of the project are the Republic of Slovenia, Ministry of Higher Education, Science and Technology and the Jožef Stefan Institute. The framework of the operation also includes the University of Ljubljana, the National Institute of Chemistry, the Institute for Electronics and Vacuum Technology and the Institute for Materials and Construction Research among others. In addition, the project is supported by the Ministry of the Economy, the National Chamber of Economy and the City of Ljubljana.

Jožef Stefan Institute
Jamova 39, 1000 Ljubljana, Slovenia
Tel.:+386 1 4773 900, Fax.:+386 1 251 93 85
WWW: http://www.ijs.si
E-mail: matjaz.gams@ijs.si
Public relations: Polona Strnad

INFORMATICA

AN INTERNATIONAL JOURNAL OF COMPUTING AND INFORMATICS

INVITATION, COOPERATION

Submissions and Refereeing

Please submit a manuscript at: http://www.informatica.si/Editors/PaperUpload.asp. At least two referees outside the author's country will examine it, and they are invited to make as many remarks as possible from typing errors to global philosophical disagreements. The chosen editor will send the author the obtained reviews. If the paper is accepted, the editor will also send an email to the managing editor. The executive board will inform the author that the paper has been accepted, and the author will send the paper to the managing editor. The paper will be published within one year of receipt of email with the text in Informatica MS Word format or Informatica LaTeX format and figures in .eps format. Style and examples of papers can be obtained from http://www.informatica.si. Opinions, news, calls for conferences, calls for papers, etc. should be sent directly to the managing editor.

QUESTIONNAIRE

□ Send Informatica free of charge

□ Yes, we subscribe

Please, complete the order form and send it to Dr. Drago Torkar, Informatica, Institut Jožef Stefan, Jamova 39, 1000 Ljubljana, Slovenia. E-mail: drago.torkar@ijs.si

Since 1977, Informatica has been a major Slovenian scientific journal of computing and informatics, including telecommunications, automation and other related areas. In its 16th year (more than twenty years ago) it became truly international, although it still remains connected to Central Europe. The basic aim of Informatica is to impose intellectual values (science, engineering) in a distributed organisation.

Informatica is a journal primarily covering intelligent systems in the European computer science, informatics and cognitive community; scientific and educational as well as technical, commercial and industrial. Its basic aim is to enhance communications between different European structures on the basis of equal rights and international refereeing. It publishes scientific papers accepted by at least two referees outside the author's country. In addition, it contains information about conferences, opinions, critical examinations of existing publications and news. Finally, major practical achievements and innovations in the computer and information industry are presented through commercial publications as well as through independent evaluations.

Editing and refereeing are distributed. Each editor can conduct the refereeing process by appointing two new referees or referees from the Board of Referees or Editorial Board. Referees should not be from the author's country. If new referees are appointed, their names will appear in the Refereeing Board.

Informatica is free of charge for major scientific, educational and governmental institutions. Others should subscribe (see the last page of Informatica).

ORDER FORM – INFORMATICA

Name: ....

Title and Profession (optional): ....

....

Home Address and Telephone (optional): ....

....

Office Address and Telephone (optional): ....

....

E-mail Address (optional): ....

Signature and Date: ....

Informatica WWW:

http://www.informatica.si/

Referees from 2008 on:

A. Abraham, S. Abraham, R. Accornero, A. Adhikari, R. Ahmad, G. Alvarez, N. Anciaux, R. Arora, I. Awan, J. Azimi, C. Badica, Z. Balogh, S. Banerjee, G. Barbier, A. Baruzzo, B. Batagelj, T. Beaubouef, N. Beaulieu, M. ter Beek, P. Bellavista, K. Bilal, S. Bishop, J. Bodlaj, M. Bohanec, D. Bolme, Z. Bonikowski, B. Bošković, M. Botta, P. Brazdil, J. Brest, J. Brichau, A. Brodnik, D. Brown, I. Bruha, M. Bruynooghe, W. Buntine, D.D. Burdescu, J. Buys, X. Cai, Y. Cai, J.C. Cano, T. Cao, J.-V. Capella-Hernández, N. Carver, M. Cavazza, R. Ceylan, A. Chebotko, I. Chekalov, J. Chen, L.-M. Cheng, G. Chiola, Y.-C. Chiou, I. Chorbev, S.R. Choudhary, S.S.M. Chow, K.R. Chowdhury, V. Christlein, W. Chu, L. Chung, M. Ciglarič, J.-N. Colin, V. Cortellessa, J. Cui, P. Cui, Z. Cui, D. Cutting, A. Cuzzocrea, V. Cvjetkovic, J. Cypryjanski, L. Čehovin, D. Čerepnalkoski, I. Čosić, G. Daniele, G. Danoy, M. Dash, S. Datt, A. Datta, M.-Y. Day, F. Debili, C.J. Debono, J. Dedič, P. Degano, A. Dekdouk, H. Demirel, B. Demoen, S. Dendamrongvit, T. Deng, A. Derezinska, J. Dezert, G. Dias, I. Dimitrovski, S. Dobrišek, Q. Dou, J. Doumen, E. Dovgan, B. Dragovich, D. Drajic, O. Drbohlav, M. Drole, J. Dujmović, O. Ebers, J. Eder, S. Elaluf-Calderwood, E. Engström, U. riza Erturk, A. Farago, C. Fei, L. Feng, Y.X. Feng, B. Filipič, I. Fister, I. Fister Jr., D. Fišer, A. Flores, V.A. Fomichov, S. Forli, A. Freitas, J. Fridrich, S. Friedman, C. Fu, X. Fu, T. Fujimoto, G. Fung, S. Gabrielli, D. Galindo, A. Gambarara, M. Gams, M. Ganzha, J. Garbajosa, R. Gennari, G. Georgeson, N. Gligorić, S. Goel, G.H. Gonnet, D.S. Goodsell, S. Gordillo, J. Gore, M. Grčar, M. Grgurović, D. Grosse, Z.-H. Guan, D. Gubiani, M. Guid, C. Guo, B. Gupta, M. Gusev, M. Hahsler, Z. Haiping, A. Hameed, C. Hamzaçebi, Q.-L. Han, H. Hanping, T. Härder, J.N. Hatzopoulos, S. Hazelhurst, K. Hempstalk, J.M.G. Hidalgo, J. Hodgson, M. Holbl, M.P. Hong, G. Howells, M. Hu, J. Hyvärinen, D. Ienco, B. Ionescu, R. Irfan, N. Jaisankar, D. Jakobović, K. Jassem, I. Jawhar, Y. Jia, T. Jin, I. Jureta, Đ. Juričić, S. K, S. Kalajdziski, Y. Kalantidis, B. Kaluža, D. Kanellopoulos, R. Kapoor, D. Karapetyan, A. Kassler, D.S. Katz, A. Kaveh, S.U. Khan, M. Khattak, V. Khomenko, E.S. Khorasani, I. Kitanovski, D. Kocev, J. Kocijan, J. Kollár, A. Kontostathis, P. Korošec, A. Koschmider, D. Košir, J. Kovač, A. Krajnc, M. Krevs, J. Krogstie, P. Krsek, M. Kubat, M. Kukar, A. Kulis, A.P.S. Kumar, H. Kwaśnicka, W.K. Lai, C.-S. Laih, K.-Y. Lam, N. Landwehr, J. Lanir, A. Lavrov, M. Layouni, G. Leban, A. Lee, Y.-C. Lee, U. Legat, A. Leonardis, G. Li, G.-Z. Li, J. Li, X. Li, X. Li, Y. Li, Y. Li, S. Lian, L. Liao, C. Lim, J.-C. Lin, H. Liu, J. Liu, P. Liu, X. Liu, X. Liu, F. Logist, S. Loskovska, H. Lu, Z. Lu, X. Luo, M. Luštrek, I.V. Lyustig, S.A. Madani, M. Mahoney, S.U.R. Malik, Y. Marinakis, D. Marinčič, J. Marques-Silva, A. Martin, D. Marwede, M. Matijašević, T. Matsui, L. McMillan, A. McPherson, A. McPherson, Z. Meng, M.C. Mihaescu, V. Milea, N. Min-Allah, E. Minisci, V. Mišić, A.-H. Mogos, P. Mohapatra, D.D. Monica, A. Montanari, A. Moroni, J. Mosegaard, M. Moškon, L. de M. Mourelle, H. Moustafa, M. Možina, M. Mrak, Y. Mu, J. Mula, D. Nagamalai, M. Di Natale, A. Navarra, P. Navrat, N. Nedjah, R. Nejabati, W. Ng, Z. Ni, E.S. Nielsen, O. Nouali, F. Novak, B. Novikov, P. Nurmi, D. Obrul, B. Oliboni, X. Pan, M. Pančur, W. Pang, G. Papa, M. Paprzycki, M. Paralič, B.-K. Park, P. Patel, T.B. Pedersen, Z. Peng, R.G. Pensa, J. Perš, D. Petcu, B. Petelin, M. Petkovšek, D. Pevec, M. Pičulin, R. Piltaver, E. Pirogova, V. Podpečan, M. Polo, V. Pomponiu, E. Popescu, D. Poshyvanyk, B. Potočnik, R.J. Povinelli, S.R.M. Prasanna, K. Pripužić, G. Puppis, H. Qian, Y. Qian, L. Qiao, C. Qin, J. Que, J.-J. Quisquater, C. Rafe, S. Rahimi, V. Rajkovič, D. Raković, J. Ramaekers, J. Ramon, R. Ravnik, Y. Reddy, W. Reimche, H. Rezankova, D. Rispoli, B. Ristevski, B. Robič, J.A. Rodriguez-Aguilar, P. Rohatgi, W. Rossak, I. Rožanc, J. Rupnik, S.B. Sadkhan, K. Saeed, M. Saeki, K.S.M. Sahari, C. Sakharwade, E. Sakkopoulos, P. Sala, M.H. Samadzadeh, J.S. Sandhu, P. Scaglioso, V. Schau, W. Schempp, J. Seberry, A. Senanayake, M. Senobari, T.C. Seong, S. Shamala, c. shi, Z. Shi, L. Shiguo, N. Shilov, Z.-E.H. Slimane, F. Smith, H. Sneed, P. Sokolowski, T. Song, A. Soppera, A. Sorniotti, M. Stajdohar, L. Stanescu, D. Strnad, X. Sun, L. Šajn, R. Šenkeřík, M.R. Šikonja, J. Šilc, I. Škrjanc, T. Štajner, B. Šter, V. Štruc, H. Takizawa, C. Talcott, N. Tomasev, D. Torkar, S. Torrente, M. Trampuš, C. Tranoris, K. Trojacanec, M. Tschierschke, F. De Turck, J. Twycross, N. Tziritas, W. Vanhoof, P. Vateekul, L.A. Vese, A. Visconti, B. Vlaovič, V. Vojisavljević, M. Vozalis, P. Vračar, V. Vranić, C.-H. Wang, H. Wang, H. Wang, H. Wang, S. Wang, X.-F. Wang, X. Wang, Y. Wang, A. Wasilewska, S. Wenzel, V. Wickramasinghe, J. Wong, S. Wrobel, K. Wrona, B. Wu, L. Xiang, Y. Xiang, D. Xiao, F. Xie, L. Xie, Z. Xing, H. Yang, X. Yang, N.Y. Yen, C. Yong-Sheng, J.J. You, G. Yu, X. Zabulis, A. Zainal, A. Zamuda, M. Zand, Z. Zhang, Z. Zhao, D. Zheng, J. Zheng, X. Zheng, Z.-H. Zhou, F. Zhuang, A. Zimmermann, M.J. Zuo, B. Zupan, M. Zuqiang, B. Žalik, J. Žižka,

Informatica

An International Journal of Computing and Informatics

Web edition of Informatica may be accessed at: http://www.informatica.si.

Subscription Information Informatica (ISSN 0350-5596) is published four times a year in Spring, Summer, Autumn, and Winter (4 issues per year) by the Slovene Society Informatika, Litostrojska cesta 54, 1000 Ljubljana, Slovenia.
The subscription rate for 2014 (Volume 38) is
– 60 EUR for institutions,
– 30 EUR for individuals, and
– 15 EUR for students
Claims for missing issues will be honored free of charge within six months after the publication date of the issue.

Typesetting: Borut Žnidar.
Printing: ABO grafika d.o.o., Ob železnici 16, 1000 Ljubljana.

Orders may be placed by email (drago.torkar@ijs.si), telephone (+386 1 477 3900) or fax (+386 1 251 93 85). The payment should be made to our bank account no.: 02083-0013014662 at NLB d.d., 1520 Ljubljana, Trg republike 2, Slovenija, IBAN no.: SI56020830013014662, SWIFT Code: LJBASI2X.

Informatica is published by Slovene Society Informatika (president Niko Schlamberger) in cooperation with the following societies (and contact persons):
Robotics Society of Slovenia (Jadran Lenarčič)
Slovene Society for Pattern Recognition (Janez Perš)
Slovenian Artificial Intelligence Society (Dunja Mladenić)
Cognitive Science Society (Urban Kordeš)
Slovenian Society of Mathematicians, Physicists and Astronomers (Andrej Likar)
Automatic Control Society of Slovenia (Sašo Blažič)
Slovenian Association of Technical and Natural Sciences / Engineering Academy of Slovenia (Vojteh Leskovšek)
ACM Slovenia (Andrej Brodnik)

Informatica is financially supported by the Slovenian research agency from the Call for co-financing of scientific periodical publications.

Informatica is surveyed by: ACM Digital Library, Citeseer, COBISS, Compendex, Computer & Information Systems Abstracts, Computer Database, Computer Science Index, Current Mathematical Publications, DBLP Computer Science Bibliography, Directory of Open Access Journals, InfoTrac OneFile, Inspec, Linguistic and Language Behaviour Abstracts, Mathematical Reviews, MatSciNet, MatSci on SilverPlatter, Scopus, Zentralblatt Math

Volume 38 Number 2 June 2014 **ISSN 0350-5596**

Informatica

An International Journal of Computing and Informatics

Informatica **38** (2014) Number 2, pp. 95–189