An Efficient Iterative Algorithm to Explainable Feature Learning

Dino Vlahek
Faculty of Electrical Engineering and Computer Science at the University of Maribor (UM FERI)
E-mail1: dino.vlahek1@um.si

Thesis Summary

Keywords: data classification, explainable artificial intelligence, feature learning

Received: April 23, 2024

This paper summarizes a doctoral thesis introducing the new iterative approach to explainable feature learning. Features are learned in three steps during each iteration: feature construction, evaluation, and selection. We demonstrated superior performances compared to the state of the art on 13 of 15 test cases and the explainability of the learned feature representation for knowledge discovery.

Povzetek: To delo povzema vsebino doktorske disertacije, v kateri predstavimo iterativni pristop k učenju razložljivih značilnic. Med vsako iteracijo se značilnice naučijo čez naslednje korake: gradnja, ocenjevanje in izbira značilnic. Na 13 od 15 testnih primerov smo demonstrirali vrhunsko zmogljivost v primerjavi s stanjem tehnike in razložljivost predstavitve naučenih značilnic za odkrivanje znanja.

1 Introduction

Supervised feature learning describes a set of techniques that enable defining augmented data representation for improved utilization of classification or regression models [1]. These methods replace traditional feature engineering tasks in a wide range of machine learning applications. Supervised feature learning methods can be divided into feature selection, dimensionality reduction, supervised dictionary learning, and deep learning. Feature selection methods select a subset of relevant features from the original feature space [3]. Such methods are limited in their accuracy as they cannot recombine features. In contrast, supervised dimensionality reduction recombines input features by mapping input samples on linear or non-linear manifolds [4]. However, significant distortions may be introduced to the data by this process as a consequence of changing distances between learning samples. Thus, resulting classification models are challenging to interpret. In addition, these approaches can only reduce the feature space's dimensionality [4]. On the other hand, supervised dictionary learning learns new feature space from the input set by recombining an arbitrary number of basic elements, called atoms, that compose a dictionary [2]. It is considered an optimization problem, where the sparsity of representation is maximized and the reconstruction errors minimized. Latter is defined as the difference between learning data and sparse representation. Dictionaries can be shared and class-specific depending on the mechanism for processing discriminatory information. Shared ones are learned from the entire data set, regardless of class labels. Using such dictionaries requires an additional classifier that significantly increases computational complexity due to the non-convex optimization problem [2]. On the other hand, class-specific dictionaries are learned for each class separately [6], enabling straightforward classification of unknown samples based on the reconstruction error introduced by such dictionaries. However, this can become computationally demanding with the increasing number of classes, while it is challenging to extract useful knowledge when the dictionary contains a large number of atoms [2]. Similar drawbacks are also noted when considering deep learning approaches. These are based on various architectures of artificial neural networks with multiple hidden layers of neurons that allow for extracting higher-level features progressively from the raw input [5]. Both linear and nonlinear functions can model neurons' activation functions, thus optimizing feature representation within the decision function. By increasing the hidden layers, artificial neural networks can approximate increasingly complex decision functions and achieve high classification accuracies. However, the presence of multiple local optima and many hyperparameters [5] also increases the training procedure's complexity, while we consider these methods as black-box function approximators [1]. In order to address above-mentioned challenges, a new method is proposed in [7] that learned interpretable features from input ones that achieved improved accuracy in comparison to the current state-of-the-art.

2 Methodology

The proposed method [7] allows for exploiting non-linear codependencies between features in order to improve an arbitrary classifier's classification performance while providing a meaningful feature representation for knowledge discovery. Each iteration consists of the following three steps: Feature construction that generates the new feature space,

feature evaluation that assesses the quality of the individual feature by a new metric that defines the feature's suitability for classification tasks, and feature selection that selects the high-quality dissimilar features using a new method based on vertex-cut. Here, we introduce two input parameters used to define the graph. The first represents the necessary level of features' quality to be included in the output feature space, and the second determines the minimal level of dissimilarity between them.

3 Results and discussion

The proposed method is extensively tested on fifteen benchmark datasets. During the sensitivity analysis, optimal values of two input parameters were identified, and the performance of five traditional classifiers was estimated on learned features. The study showed that the learned features statistically significantly improved the classification accuracy of all tested classifiers, while the random forest classifier achieved the best results. As demonstrated by experiments, the proposed method achieved or exceeded the classification accuracy of six state-of-the-art in all test cases. The correctness of learned features interpretation on a well-studied dataset was also demonstrated.

The proposed method is used in many research applications, ranging from pure research to industrial and scientific projects. We plan to extend the proposed method application to regression with a new feature evaluation metric for the suitability of features for regression tasks.

References

[1] Y. Bengio, A. Courville, and P. Vincent. Representation Learning: A Review and New Perspectives. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 35(8):1798–1828, aug 2013. doi: https://doi.org/10.1109/TPAMI.2013.50.

[2] Mehrdad J. Gangeh, Ahmed K. Farahat, Ali Ghodsi, and Mohamed S. Kamel. Supervised dictionary learning and sparse representation-a review. *ArXiv*, abs/1502.05928, 2015. doi: https://doi.org/10.48550/arXiv.1502.05928.

[3] Huan Liu and Hiroshi Motoda. *Feature Extraction, Construction and Selection: A Data Mining Perspective*. Kluwer Academic Publishers, Norwell, MA, USA, 1998.

[4] Yunqian Ma and Yun Fu. *Manifold Learning Theory and Applications*. CRC Press, Inc., USA, 1st edition, 2011.

[5] Michael A. Nielsen. *Neural Networks and Deep Learning*, page 216. Determination Press, 2018.

[6] W. Tang, A. Panahi, H. Krim, and L. Dai. Analysis dictionary learning based classification: Structure for robustness. *IEEE Transactions on Image Processing*, 28(12):6035–6046, Dec 2019. doi: http://doi.org//10.1109/TIP.2019.2919409.

[7] Dino Vlahek. Učinkovit iterativni algoritem učenja razložljivih značilnic za izboljšano klasifikacijo. *PhD thesis, UM FERI*, 2024.