Information Visualization using Machine Learning

Gregor Leban
Institut Jožef Stefan, Jamova cesta 39, Slovenia
E-mail: gregor.leban@ijs.si

Thesis Summary

Keywords: information visualization, machine learning, radviz, parallel coordinates, mosaic plot

Received: February 20, 2013

Data visualization is an important tool for discovering patterns in the data. Finding interesting visualizations can be however a difficult task if there are many possible ways to visualize the data. In this paper we present the VizRank method that can estimate visualization interestingness. The method can be applied on a number of visualization techniques and can automatically identify the most interesting data visualizations.

Povzetek: Predstavljena metoda VizRank omogoča avtomatsko ocenjevanje zanimivosti različnih vizualizacij podatkov in posledično identifikacijo najzanimivejših prikazov. Metodo je mogoče uporabiti na poljubni metodi s točkovnim prikazom podatkov, na metodi paralelnih koordinat ter na mozaičnih diagramih.

1 Introduction

Data visualization has a great potential for extracting knowledge from data. Visualizing the right set of features can clearly identify interesting patterns. However, not all data projections are equally interesting and the task of the data analyst is to find the most insightful ones. In case of supervised learning, we are looking for those visualizations that show clear class separation. Finding such visualizations (if they exist), can be very challenging especially when there are many possible ways to visualize the data.

In order to make the task easier we developed a method called VizRank that can be used to automatically identify the most interesting visualizations of a dataset. The method was developed to be used on all point-based visualization methods such as scatterplot, radviz, polyviz and linear projections. We later extended it also for use on parallel coordinates and mosaic plots.

In the paper we will mention two less commonly known visualization methods - radviz[1] and parallel coordinates[2]. In radviz the visualized features are represented as dots distributed along the circle. For each data example, each dot (feature) is attracting the example with a force corresponding to the value of that feature - the greater the value, the greater the attraction. The example is displayed where the sum of forces equals 0. In parallel coordinates, the axes for the visualized features are displayed parallel to each other. Each example is shown as a series of lines that intersect each axis at the point that corresponds to the value of the feature.

2 VizRank

VizRank[4] identifies the most interesting visualizations by repeating the following steps. First, a method for generating different feature subsets is used to select a set of features to be visualized and evaluated. Given the features, the positions of data points in the projection are then determined based on the chosen visualization method. A new dataset is then constructed consisting of only the x and y data point positions and their labels. A machine learning algorithm is then used on this dataset to evaluate the quality of class separation. The computed accuracy of the algorithm is used as the score of the interestingness of the projection.

Method for generating different feature subsets. The space of possible visualizations is commonly too large to evaluate all possible visualizations. To identify interesting projections fast and by checking only a small subset of possible projections we developed different heuristic methods. The one that performs best starts by ranking individual features using a feature selection method such as ReliefF[3]. From the ranked list of features we then choose the desired number of features using the gamma probability distribution. With this sampling method, the higher ranked features are selected more often. The intuition for this approach is that the features that are better at class separation will more likely generate interesting visualizations and should be tested more often than features that are worse.

Learning algorithm. Humans are able to detect arbitrarily shaped class boundaries in the visualizations. In order to best mimic humans we decided to use k-NN as the learning method. We experimented with different scoring functions such as classification accuracy and Brier score. At the end

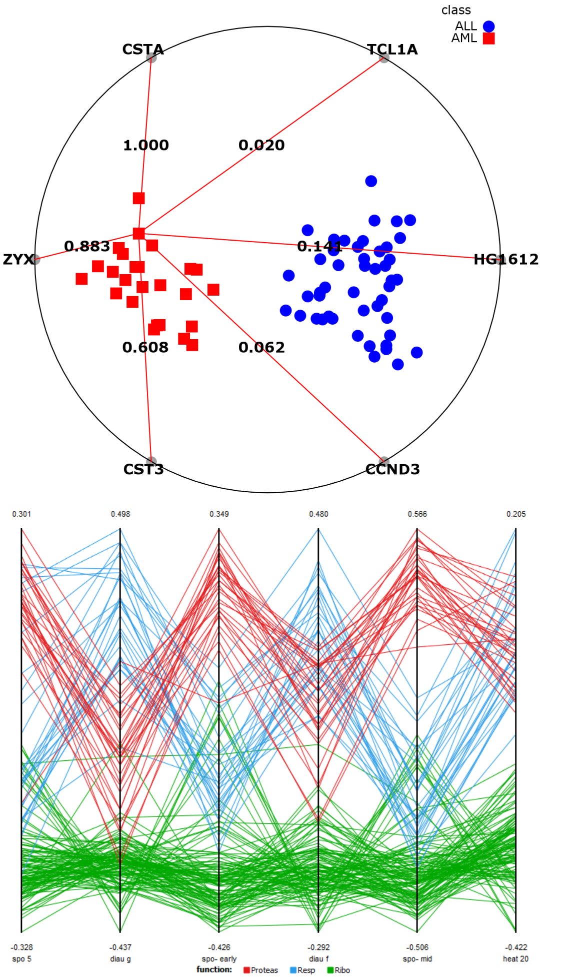

Figure 1: Two visualizations identified by VizRank - radviz visualization of *leukemia* dataset (top) and parallel coordinates plot of *yeast* dataset (bottom).

used the average probability assigned to the correct class which is defined as

$\frac{1}{N}\sum_{i=1}^{N} P(y_i|x_i)$

We chose this method since our experiments indicated that it produces very refined and human-like ranking of projections.

Some uses of best ranked projections. VizRank produces as a result a ranked list of projections. One possible use of this list is to perform feature scoring. In this case, the features are scored based on how often they appear in the top ranked projections. Instead of myopic measures that score each feature independently of the others, this measure can also identify features that are important when combined with other features. The list of projections can also be used for outlier detection. Frequently in top ranked projections some points lie outside of their main cluster of points. To understand if the example is really an outlier we can visualize the class prediction of the point in several top ranked projections.

Agreement with human ranking. Our base assumption in VizRank is that projections with high prediction accuracy are most insteresting for the data analysts. To evaluate how well does ranking obtained using VizRank actually correspond to ranking done by humans we performed an experiment in which 30 people ranked 20 pairs of projections. The obtained correlation between VizRank and human ranking was 0.78. To test the influence of the learning algorithm we also ranked projections using SVM and decision trees instead of k-NN. Using SVM the correlation fell to 0.58, while using decision trees it dropped to only 0.28. Results confirm that k-NN is the most appropriate of the tested methods and that ranking results highly correlate with human ranking.

Use on other visualization methods. The method, as presented, can be applied on any point based visualization method. We extended VizRank also to parallel coordinates and mosaic plots by identifying the desired properties that interesting visualizations have. In case of parallel coordinates, for example, examples from each class should be drawn under similar angle. This reduces clutter caused by intersecting lines and allows the detection of a regular pattern. By defining a corresponding optimization function we were then able to identify visualizations as the one in Figure 1.

3 Conclusion

We developed and presented the VizRank method that can automatically evaluate interestingness of different visualizations of labeled data. It is most valuable when the analysed data contains hundreds or thousands of features which makes manual search for interesting visualizations impractical. Empirical results confirm that k-NN is the most appropriate of the learning algorithms and that ranking of projections obtained with VizRank highly correlates with human ranking. The method can be applied on any point-based method as well as on parallel coordinates and mosaic plots.

References

[1] G. Grinstein, M. Trutschl, and U. Cvek. High-dimensional visualizations. *Proceedings of the Visual Data Mining Workshop, KDD*, 2001.

[2] Alfred Inselberg. *Parallel coordinates: visual multidimensional geometry and its applications*. Springer, 2009.

[3] I. Kononenko and E. Simec. Induction of decision trees using relieff. In *Mathematical and statistical methods in artificial intelligence*. Springer Verlag, 1995.

[4] G. Leban, I. Bratko, U. Petrovic, T. Curk, and B. Zupan. Vizrank: finding informative data projections in functional genomics by machine learning. *Bioinformatics*, 21(3):413–414, 2005.