Erschienen in: EuroVis 2016 - Posters / Isenberg, Tobias; Sadlo, Filip (Hrsg.). - The Eurographics Association, 2016. - ISBN 978-3-03868-015-4
https://dx.doi.org/10.2312/eurp.20161130

When individual data points matter: interactively analysing classification landscapes

Bruno Schneider, Sebastian Mittelstädt and Daniel Keim

University of Konstanz, Germany

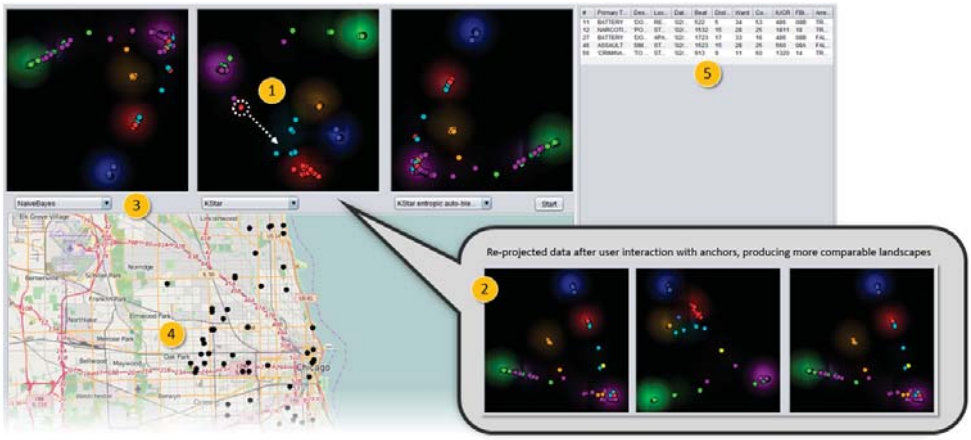

Figure 1: *Our system provides visual model comparison by (1) showing side-by-side the classification landscapes of different models in linked panels after applying MDS projection on the probabilities estimates for each class of each data instance, (2) enabling interactive anchor-points selection, dragging and data reprojection for producing more comparable landscapes, (3) allowing the selection of other previously evaluated models, (4) letting the user collaborate with domain knowledge through direct selection of geo-referenced events on a map, and also (5) giving detailed information of each data instance in a text table.*

Abstract

The selection of classification models among several options with similar accuracy cannot be done through purely automated methods, and especially in scenarios in which the cost of misclassified instances is crucial, such as criminal intelligence analysis. To tackle this problem and illustrate our ideas, we developed a prototype for the visualization and comparison of classification landscapes. In our system, the same data is given to different classification models. Classification landscapes are shown in the scatter plots, together with their geographical location on a map and detailed textual description for each data record. To enhance model comparison, we implemented interactive anchor-points selection in classification landscapes. Using those anchors, the user can manipulate and reproject the model results in order to get more comparable classification landscapes. We provided a use case with crime data, for crime intelligence analysis.

Categories and Subject Descriptors (according to ACM CCS): I.5.2 [Pattern Recognition]: Design Methodology—Classifier design and evaluation

1. Introduction

In some application scenarios of data analysis, the cost of misclassified or misinterpreted data instances is crucial. In crime intelligence analysis, for example, wrong classified data records could influence the life and reputation of persons critically. We argue that the real accuracy of a model cannot be estimated without domain and expert knowledge. The decision making processes require an understanding of model and data at the same time. Automatic methods oppose dangers in the decision making process since every dimensional-

ity reduction (classification and projection) or aggregation hides important information for detailed analysis.

Interactive machine learning provides methods to build accurate models by integrating expert knowledge. Typically, performance measures evaluate the model while interactive refinement focuses on optimizing these measures or classification borders. Therefore, it is possible to build and refine several models with appropriate accuracy based on *global* measures. The emergent question is: which model to select if there are multiple models with similar accuracy?

In our target application scenario, however, *local* patterns matter because decisions are based on individual records which are not captured in *global* measures and classification borders. Therefore, we contribute with a classification landscape visualization based on multidimensional scaling (MDS) that allows *global* model comparison and analysis of *local* patterns. We extended an existing MDS algorithm [Gro09] by interactive anchor-points selection that aims to stabilize different projection results. The anchored projection thereby allows to experience changes of models and parameters and, thus, supports experts in understanding and comparing models.

2. Related Work

Visual techniques for the evaluation of classification results were presented previously [DA08, KS12, RD00] and further for model comparison during the process of model building [CCWH08]. Lately, Alsallakh et al. [AHH*14] visually analyze class probabilities estimates. Kapoor et al. [KLTH10] apply confusion matrices for interactive optimization of model. The aforementioned works are rather focussed on the aggregates of probabilites or global performance measures and do not allow local pattern analysis. Differently, Migut et al. [MWV15] apply scatter plots on non-aggregated data. However, the attributes of the data were plotted directly, which prevents the user to estimate global classification borders and local classifications patterns.

3. System Archictecture

We developed a prototype for the comparison of classification landscapes produced by different machine learning classifiers, and after the process of model building. The visual components and our workflow were designed to support analysis tasks in which it is important to compare classification borders between different models with similar global accuracy, while preserving individual data instances. Also, in our system the user can load up to three previously built classification models for visual exploration of the outputs.

The classification results are shown in scatter plots, one for each model, after applying MDS projection on the input data. Each class is represented by a different color and high densities of records of the same class reveal position of classes in the landscape. Regarding the data projection, we applied multidimensional scaling on the probabilities estimates of each data instance for each predicted class, instead of projecting all the original data attributes. By projecting the data with fewer dimensions, we minimized the problems that arise with the curse of dimensionality.

All visualization panels are interconnected by linking and brushing, enabling top-down analysis from global classification borders to local crime patterns in their geographic context. The crime events can also be analyzed in a geo-referenced context, together with a text table providing details for each crime event.

The global comparison of classification landscapes can be difficult since the class positions and borders may appear in different locations in different scatter plots. Therefore, we implemented interactive anchor-points selection, in which the user can select single records, drag, and anchor them in the scatter plots and, thus, within the projection. Then, when the user drags a data point the MDS projection is recomputed without moving the anchored records in the stress minimization. Thus, the anchored records will appear at the same position of all plots. Additionally, all other points are moved accordingly in the projection, e.g., dragging all similar records towards anchors and thereby also class positions. Since all classification landscapes share the same anchor points, global comparison can be performed with less cognitive load.

4. Use case: Interactive analysis of crime classification data

We have chosen to work with a crime dataset from the city of Chicago, U.S. [Dep]. This choice fits our proposal focused on domains where it is important to keep track of individual events, due to the sensibility of related issues and resulting high costs of wrong predictions. The attributes selected for prediction were the most frequent six types of crimes from the dataset, in a way similar to the competition organized in [Kag15].

We trained seven different classification models. In this group, we have one model based on Neural Networks, a Support Vector Machine model, a KStar instance-based classifier, Decision Trees and also a K-nearest neighbours model. Some of them gave us very good accuracy for our classification task (results ranging from 72 to 99% of accuracy).

Regarding findings, our tool achieved initial good results in a scenario where we had the same classification model with 3 different parameter settings. Then, by manipulating the data-points as shown in Figure 1, we generated much more comparable shapes among plots after user interaction and data manipulation than the initially generated automatic projections by the system. To choose which point to select as an anchor and drag it, we prioritized points that were classified differently by one single model and moved it into the same direction that this point was classified by the other models.

5. Conclusion and Future Work

We see our work as an effort for a better understanding of individually classified data instances, with an important extension for geo-referenced data. The proposed framework and the way we applied the data projections with interactive anchor-points selection could be incorporated into more extensive predictive modeling workflows. Also, the interactive anchor-points helped in providing more comparable classification landscapes, through direct user manipulation supported by contextual information.

Despite our efforts on bringing the most of contextual information as possible to the analysis of the results produced by different classifiers, this work can be extended towards the field of "Interactive Machine Learning" [PTH13], where the users can improve the performance of different classifiers through visual means.

References

[AHH*14] ALSALLAKH B., HANBURY A., HAUSER H., MIKSCH S., RAUBER A.: Visual methods for analyzing probabilistic classification data. *Visualization and Computer Graphics, IEEE Transactions on 20*, 12 (2014), 1703–1712. 2

[CCWH08] CARAGEA D., COOK D., WICKHAM H., HONAVAR V.: Visual methods for examining svm classifiers. In *Visual Data Mining*. Springer, 2008, pp. 136–153. 2

[DA08] DIRI B., ALBAYRAK S.: Visualization and analysis of classifiers performance in multi-class medical data. *Expert Systems with Applications 34*, 1 (2008), 628–634. 2

[Dep] DEPARTMENT C. P.: Crimes - 2001 to present. URL: http://data.cityofchicago.org/Public-Safety/Crimes-2001-to-present/ijzp-q8t2. 2

[Gro09] GROUP A.: Mdsj: Java library for multidimensional scaling (version 0.2), University of Konstanz, 2009. Available at http://www.inf.uni-konstanz.de/algo/software/mdsj/. 2

[Kag15] KAGGLE: San francisco crime classification, 2015. URL: https://www.kaggle.com/c/sf-crime. 2

[KLTH10] KAPOOR A., LEE B., TAN D., HORVITZ E.: Interactive optimization for steering machine classification. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (2010), ACM, pp. 1343–1352. 2

[KS12] KIENREICH W., SEIFERT C.: Visual exploration of feature-class matrices for classification problems. In *International Workshop on Visual Analytics (EuroVA)* (2012), pp. 37–41. 2

[MWV15] MIGUT M., WORRING M., VEENMAN C.: Visualizing multi-dimensional decision boundaries in 2d. *Data Mining and Knowledge Discovery 29*, 1 (2015), 273–295. 2

[PTH13] PORTER R., THEILER J., HUSH D.: Interactive machine learning in data exploitation. *Computing in Science & Engineering 15*, 5 (2013), 12–20. 2

[RD00] RHEINGANS P., DESJARDINS M.: Visualizing high-dimensional predictive model quality. In *Proceedings of the conference on Visualization'00* (2000), IEEE Computer Society Press, pp. 493–496. 2