Master Thesis

Karlsruher Institut für Technologie

Institut für Programmstrukturen und Datenorganisation (IPD)
Lehrstuhl für Systeme der Informationsverwaltung, Prof. Böhm

Master Thesis

Evaluating Methods for Imbalanced Learning

Classification algorithms in machine learning often have to deal with imbalanced data. Consider the scenario of fraud detection for credit card transactions: There are many more genuine transactions than fraudulent ones. For example, assume that only 1% of the transactions are fraudulent. Now suppose that we want to build a classifier to predict if a transaction is fraudulent or honest. If the classifier and the subsequent evaluation focus on accuracy, you can easily achieve 99% correct predictions simply by always predicting "honest" – without ever detecting a fraudulent transaction.

Appropriately handling imbalanced data in classification – in the following called "imbalanced learning" – is an important topic in machine learning. Researchers have developed a large variety of approaches over the years, e.g., sampling the training set, weighting instances or classes, treating the problem as outlier detection etc. However, there is a lack of research regarding truly broad comparative studies. Many approaches are only evaluated against a few closely related competitors. Datasets in evaluations are sometimes only from one domain or are purely synthetic. Furthermore, the imbalanced learning approaches might have a varying impact on different classification algorithms and different classification evaluation metrics. It is unclear which approach one should use under which circumstances.

The goal of this thesis is to create a framework to evaluate imbalanced learning methods systematically. The following questions are particularly interesting:

- Which characteristics of a dataset influence the performance of imbalanced learning?
- Under which settings should you choose a certain imbalanced learning method?
- Is there an imbalanced learning method consistently outperforming others?
- How good are rather simple imbalanced learning methods compared to sophisticated ones?
- Which influence does the choice of the classification evaluation metric have when assessing the performance of imbalanced learning approaches?

The following steps are part of your thesis:

- Review literature about classifying imbalanced data.
- Design and implement a framework to evaluate classification performance with multiple imbalanced learning approaches, datasets, classifiers and evaluation metrics. You can use existing machine learning libraries for individual components of your framework, as well as datasets from public repositories like the UCI Machine Learning Repository.
- Experimental evaluation of your approach, drawing conclusions from a large set of results. You can use the server infrastructure of our chair to run experiments.

During your work on this thesis, you will acquire practical knowledge about state-of-the-art machine learning libraries. You will get familiar with various imbalanced learning approaches and gain an understanding of their usefulness as well as limitations for different kinds of datasets. You will gain experience in running and evaluating large-scale scientific experiments.

You can write the thesis in English or German. Prior experience with classification in any programming language (e.g. Python, R) is beneficial, but not necessary.

Contact

Jakob Bach | jakob.bach@kit.edu | +49 721 608-47339 | Room: 351
Am Fasanengarten 5 | 76131 Karlsruhe | Building: 50.34