Data Summarisation for Imbalanced Datasets

Supervisor: Aleksei Triastcyn

Keywords: machine learning, data summarisation, imbalanced datasets, asymmetric loss functions

Description

In many real-world classification problems the number of instances is not balanced between classes, which makes these problems extremely difficult for most machine learning algorithms. Examples of such data include network intrusion, fault and fraud detection, medical anomaly detection, and so on. There has been a great variety of research on working with imbalanced data, but it still poses a serious challenge. The goal of this project is to try to overcome difficulties of data summarisation and consecutive learning from these summaries for such datasets.

During the course of the project, a student will implement a data summarisation method based on submodular functions [1, 2], as well as a number of simple baseline methods to compare with. Then, evaluate and compare these techniques on several imbalanced datasets with asymmetric loss functions, such that, for example, the cost of false negatives is much higher than the cost of false positives. Finally, determine if there is an advantage in using submodular data summarisation in such a scenario.

Project Objectives

- Get familiar with the concept of data summarisation with submodular functions.
- Implement a summarisation algorithm.
- Apply data summarisation to imbalanced datasets and compare with baseline methods.

References

[1] A. Krause, D. Golovin, Submodular function maximization. In *Tractability: Practical Approaches to Hard Problems*, 2012.

[2] H. Zhuang, R. Rahman, X. Hu, T. Guo, P. Hui, K. Aberer, Data Summarization with Social Contexts. In *CIKM*, 2016.