# DOCTORAL THESIS

## Advances in imbalanced data learning

Lu, Yang

*Date of Award:*
2019

# Abstract

With the increasing availability of large amount of data in a wide range of applications, no matter for industry or academia, it becomes crucial to understand the nature of complex raw data, in order to gain more values from data engineering. Although many problems have been successfully solved by some mature machine learning techniques, the problem of learning from imbalanced data continues to be one of the challenges in the field of data engineering and machine learning, which attracted growing attention in recent years due to its complexity. In this thesis, we focus on four aspects of imbalanced data learning and propose solutions to the key problems.

The first aspect is about ensemble methods for imbalanced data classification. Ensemble methods, e.g. bagging and boosting, have the advantages to cure imbalanced data by integrated with sampling methods. However, there are still problems in the integration. One problem is that undersampling and oversampling are complementary each other and the sampling ratio is crucial to the classification performance. This thesis introduces a new method HSBagging which is based on bagging with hybrid sampling. Experiments show that HSBagging outperforms other state-of-the-art bagging method on imbalanced data. Another problem is about the integration of boosting and sampling for imbalanced data classification. The classifier weights of existing AdaBoost-based methods are inconsistent with the objective of class imbalance classification. In this thesis, we propose a novel boosting optimization framework GOBoost. This framework can be applied to any boosting-based method for class imbalance classification by simply replacing the calculation of clas-

sifier weights. Experiments show that the GOBoost-based methods significantly outperform the corresponding boosting-based methods.

The second aspect is about online learning for imbalanced data stream with concept drift. In the online learning scenario, if the data stream is imbalanced, it will be difficult to detect concept drifts and adapt the online learner to them. The ensemble classifier weights are hard to adjust to achieve the balance between the stability and adaptability. Besides, the classier built on samples in size-fixed chunk, which may be highly imbalanced, is unstable in the ensemble. In this thesis, we propose Adaptive Chunk-based Dynamic Weighted Majority (ACDWM) to dynamically weigh the individual classifiers according to their performance on the current data chunk. Meanwhile, the chunk size is adaptively selected by statistical hypothesis tests. Experiments on both synthetic and real datasets with concept drift show that ACDWM outperforms both of the state-of-the-art chunk-based and online methods.

In addition to imbalanced data classification, the third aspect is about clustering on imbalanced data. This thesis studies the key problem of imbalanced data clustering called uniform effect within the k-means-type framework, where the clustering results tend to be balanced. Thus, this thesis introduces a new method called Self-adaptive Multi-prototype-based Competitive Learning (SMCL) for imbalanced clusters. It uses multiple subclusters to represent each cluster with an automatic adjustment of the number of subclusters. Then, the subclusters are merged into the final clusters based on a novel separation measure. Experimental results show the efficacy of SMCL for imbalanced clusters and the superiorities against its competitors.

Rather than a specific algorithm for imbalanced data learning, the final aspect is about a measure of class imbalanced dataset for classification. Recent studies have shown that imbalance ratio is not the only cause of the performance loss of a classifier in imbalanced data classification. To the best of our knowledge, there is no any measurement about the extent of influence of class imbalance on the

classification performance of imbalanced data. Accordingly, this thesis proposes a data measure called Bayes Imbalance Impact Index ($BI^3$) to reflect the extent of influence purely by the factor of imbalance for the whole dataset. As a result, we can therefore use $BI^3$ to judge whether it is worth using imbalance recovery methods like sampling or cost-sensitive methods to recover the performance loss of a classifier. The experiments show that $BI^3$ is highly consistent with the improvement of F1 score made by the imbalance recovery methods on both synthetic and real benchmark datasets.

In summary, the major contributions of this thesis is listed as follows.

- Two ensemble frameworks for imbalanced data classification are proposed for sampling rate selection and boosting weight optimization, respectively.
- A chunk-based online learning algorithm is proposed to dynamically adjust the ensemble classifiers and select the chunk size for imbalanced data stream with concept drift.
- A multi-prototype competitive learning algorithm is proposed for clustering on imbalanced data.
- A measure of imbalanced data is proposed to evaluate how the classification performance of a dataset is influenced by the factor of imbalance.

**Keywords:** Imbalanced data learning, ensemble methods, online learning, imbalanced data stream, concept drift, imbalanced data clustering, imbalance impact.

# Table of Contents