

DOCTORAL THESIS

Advances in imbalanced data learning

Lu, Yang

Date of Award:
2019

[Link to publication](#)

General rights

Copyright and intellectual property rights for the publications made accessible in HKBU Scholars are retained by the authors and/or other copyright owners. In addition to the restrictions prescribed by the Copyright Ordinance of Hong Kong, all users and readers must also observe the following terms of use:

- Users may download and print one copy of any publication from HKBU Scholars for the purpose of private study or research
- Users cannot further distribute the material or use it for any profit-making activity or commercial gain
- To share publications in HKBU Scholars with others, users are welcome to freely distribute the permanent URL assigned to the publication

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 classifier 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

Declaration	i
Abstract	ii
Acknowledgements	v
Table of Contents	vii
List of Tables	xi
List of Figures	xvi
Chapter 1 Introduction	1
1.1 Learning From Imbalanced Data	1
1.2 Application	4
1.3 Measurements	6
1.4 Review of Methods for Imbalanced Data Learning	7
1.4.1 Sampling Methods	8
1.4.2 Cost-sensitive Methods	9
1.4.3 Ensemble Methods	11
1.5 Contribution	14
1.6 Organization	15
Chapter 2 Ensemble Methods for Imbalanced Data Classification	17
2.1 Hybrid Sampling with Bagging for Class Imbalance Learning	17

2.1.1	Introduction	18
2.1.2	Related Work	20
2.1.3	The Proposed Method	21
2.1.4	Experiments	23
2.2	GOBoost: G-mean Optimized Boosting Framework for Class Imbalance Learning	31
2.2.1	Introduction	31
2.2.2	Related Work	33
2.2.3	Optimization Objective of AdaBoost	35
2.2.4	GOBoost Framework	36
2.2.5	Experiments	42
2.3	Summary	53
Chapter 3 Online Learning for Imbalanced Data Stream with Concept Drift		
	Drift	55
3.1	Introduction	56
3.2	Related Work	60
3.2.1	Online Methods	61
3.2.2	Chunk-based Methods	61
3.2.3	Window Selection Methods for Drift Detection	62
3.3	Proposed Method	63
3.3.1	General Framework	63
3.3.2	Chunk Training	66
3.3.3	Adaptive Chunk Size Selection	70
3.4	Experiments	76
3.4.1	Experiment Settings	76
3.4.2	Datasets	79
3.4.3	Experimental Results	81
3.5	Summary	97

Chapter 4 Clustering on Imbalanced Data	98
4.1 Introduction	99
4.2 Overview of Related Work	103
4.2.1 k -means-type Competitive Learning	103
4.2.2 Nonlinear Clustering	106
4.2.3 Imbalance Clustering	106
4.3 The Proposed Method	107
4.3.1 Selection of Number of Prototypes	109
4.3.2 Subcluster Grouping with Model Selection	114
4.4 Experimental Results	119
4.4.1 Datasets and Compared Methods	119
4.4.2 Evaluation Metrics	122
4.4.3 Results and Discussion	122
4.4.4 Parameter Sensitivity	130
4.4.5 Running Time Analysis	132
4.5 Summary	134
Chapter 5 Measure of Class Imbalanced Dataset for Classification	
Problem	135
5.1 Introduction	136
5.2 Related Work	139
5.2.1 Data Complexity for Imbalanced Data Classification	139
5.3 Proposed Method	141
5.3.1 Guidance of Usage	149
5.4 Experiments	150
5.4.1 Synthetic Data	152
5.4.2 Real Benchmark Data	161
5.5 Summary	169
Chapter 6 Conclusion	170

Bibliography	174
List of Publications	195
CURRICULUM VITAE	197