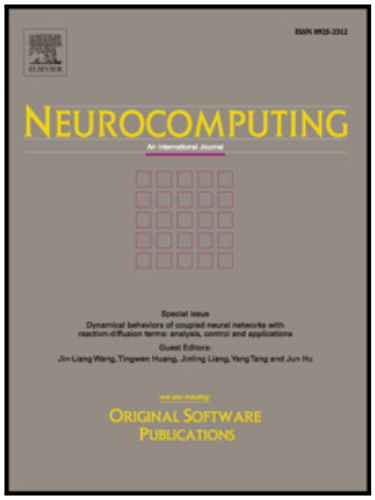
Communicated by Dr Xiaofeng Zhu

Accepted Manuscript



The Gradual Resampling Ensemble for mining imbalanced data streams with concept drift

Siqi Ren, Bo Liao, Wen Zhu, Zeng Li, Wei Liu, Keqin Li

PII: S0925-2312(18)30096-1
DOI: 10.1016/j.neucom.2018.01.063
Reference: NEUCOM 19265

To appear in: *Neurocomputing*

Received date: 9 March 2017
Revised date: 10 November 2017
Accepted date: 24 January 2018

Please cite this article as: Siqi Ren, Bo Liao, Wen Zhu, Zeng Li, Wei Liu, Keqin Li, The Gradual Resampling Ensemble for mining imbalanced data streams with concept drift, *Neurocomputing* (2018), doi: 10.1016/j.neucom.2018.01.063

The Gradual Resampling Ensemble for mining imbalanced data streams with concept drift

Siqi Ren[a], Bo Liao[a,*], Wen Zhu[a], Zeng Li[b], Wei Liu[a], Keqin Li[a,c]

[a]*College of Information Science and Engineering, Hunan University, Changsha 410082, Hunan, China*
[b]*School of Computer Science and Technology, University of Science and Technology of China, Hefei 230027, Anhui, China*
[c]*Department of Computer Science, State University of New York, New Paltz, New York 12561, USA*

Abstract

Knowledge extraction from data streams has received increasing interest in recent years. However, most of the existing studies assume that the class distribution of data streams is relatively balanced. The reaction of concept drifts is more difficult if a data stream is class imbalanced. Current oversampling methods generally selectively absorb the previously received minority examples into the current minority set by evaluating similarities of past minority examples and the current minority set. However, the similarity evaluation is easily affected by data difficulty factors. Meanwhile, these oversampling techniques have ignored the majority class distribution, thus risking class overlapping.

To overcome these issues, we propose an ensemble classifier called Gradual Resampling Ensemble (GRE). GRE could handle data streams which exhibit concept drifts and class imbalance. On the one hand, a selectively resampling method, where drifting data can be avoidable, is applied to select a part of previous minority examples for amplifying the current minority set. The disjuncts can be discovered by the DBSCAN clustering, and thus the influences of small disjuncts and outliers on the similarity evaluation can be avoidable. Only those minority examples with low probability of overlapping with the current majority set can be selected for resampling the current minority set. On the other hand, previous component classifiers are updated using latest instances. Thus, the ensemble could quickly adapt to a new condition, regardless types of concept drifts. Through the gradual oversampling of previous chunks using the current minority events, the class distribution of past chunks can be balanced. Favorable results in comparison to other algorithms suggest that GRE can maintain good performance on minority class, without sacrificing majority class performance.

Keywords:
Concept drift, Data stream mining, Ensemble classifier, Class imbalance .

*Corresponding author
Email addresses: siqirenzl@163.com (Siqi Ren), dragonbw@163.com (Bo Liao), syzhuwen@163.com (Wen Zhu), lizeng@mail.ustc.edu.cn (Zeng Li), lw2001184@163.com (Wei Liu), lik@newpaltz.edu (Keqin Li)