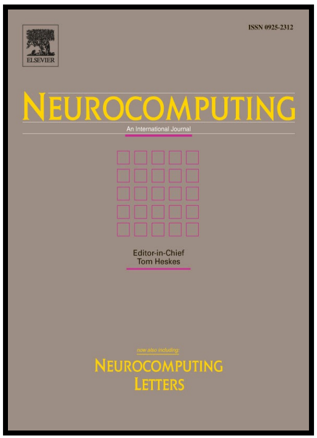

www.elsevier.com/locate/neucom

A New Sampling Method for Classifying Imbalanced Data Based on Support Vector Machine Ensemble

Chuanxia Jian, Jian Gao, Yinhui Ao

PII: S0925-2312(16)00184-3
DOI: http://dx.doi.org/10.1016/j.neucom.2016.02.006
Reference: NEUCOM16729

To appear in: *Neurocomputing*

Received date: 11 June 2015
Revised date: 23 December 2015
Accepted date: 4 February 2016

Cite this article as: Chuanxia Jian, Jian Gao and Yinhui Ao, A New Sampling Method for Classifying Imbalanced Data Based on Support Vector Machin Ensemble, *Neurocomputing,* http://dx.doi.org/10.1016/j.neucom.2016.02.006

A New Sampling Method for Classifying Imbalanced Data Based on Support Vector Machine Ensemble

Chuanxia Jian, Jian Gao, Yinhui Ao

(Key Laboratory of Mechanical Equipment Manufacturing & Control Technology of Ministry of Education，School of Electromechanical Engineering，Guangdong University of Technology，Guangzhou 510006, P.R. China)

Abstract：The insufficient information from the minority examples can not exactly represent the inherent structure of the dataset, which leads to a low prediction accuracy of the minority through the existing classification methods. The over- and under-sampling methods help to increase the prediction accuracy of the minority. However, the two methods either lose important information or add trivial information for classification, so as to affect the prediction accuracy of the minority. Therefore, a new different contribution sampling method (DCS) based on the contributions of the support vectors (SVs) and the nonsupport vectors (NSVs) to classification is proposed in this paper. The proposed DCS method applies different sampling methods for the SVs and the NSVs and uses the biased support vector machine (B-SVM) method to identify the SVs and the NSVs of an imbalanced data. Moreover, the synthetic minority over-sampling technique (SMOTE) and the random under-sampling technique (RUS) is used in the proposed method to re-sample the SVs in the minority and the NSVs in the majority, respectively. Examples are labeled by the ensemble of support vector machine (SVM$_{en}$). Experiments are carried out on the imbalanced dataset which is selected from UCI, AVU06a, Statlog, DP01a, JP98a and CWH03a repositories. Experimental results show that for the imbalanced datasets, the proposed DCS method achieves a better performance in the aspects of receiver operating characteristic (ROC) curve than other methods. The proposed DCS method improves 20.80%, 5.97%, 8.66% and 9.35% in terms of the geometric mean prediction accuracy G_{mean} as compared with that achieved by using the NS, the US, the SMOTE and the ROS, respectively.

Keywords: Imbalanced data, sampling, support vector machine.

1. INTRODUCTION

In the areas of mechanical engineering, financial systems, information and medical science, skew data appears frequently in their datasets[1-4]. The models created by the imbalanced data will result in a low prediction accuracy of the minority. In the dataset, suppose that the number of data labeled with -1(the negative class) is far more than those labeled with +1 (the positive class). The importance of the minority sometimes outperforms the majority. For instance, the occurrence of defect examples (the minority examples) is much less than the normal ones, and it is very important to find them from all products[5, 6].