Information Sciences 178 (2008) 3214–3227

Contents lists available at ScienceDirect

Information Sciences

journal homepage: www.elsevier.com/locate/ins

An information granulation based data mining approach for classifying imbalanced data

Mu-Chen Chen [a,*], Long-Sheng Chen [b], Chun-Chin Hsu [c], Wei-Rong Zeng [d]

[a] *Institute of Traffic and Transportation, National Chiao Tung University, 4F, 118, Section 1, Chung-Hsiao W. Road, Taipei 10012, Taiwan*
[b] *Department of Information Management, Chaoyang University of Technology, 168, Jifong E. Road, Wufong Township, Taichung County 41349, Taiwan*
[c] *Department of Industrial Engineering and Management, Chaoyang University of Technology, 168, Jifong E. Road, Wufong Township, Taichung County 41349, Taiwan*
[d] *Information Management Department, Entie Commercial Bank, Taipei, Taiwan*

ARTICLE INFO

Keywords:
Information granulation
Granular computing
Data mining
Latent semantic indexing
Imbalanced data
Feed-forward neural network

ABSTRACT

Recently, the class imbalance problem has attracted much attention from researchers in the field of data mining. When learning from imbalanced data in which most examples are labeled as one class and only few belong to another class, traditional data mining approaches do not have a good ability to predict the crucial minority instances. Unfortunately, many real world data sets like health examination, inspection, credit fraud detection, spam identification and text mining all are faced with this situation. In this study, we present a novel model called the “Information Granulation Based Data Mining Approach” to tackle this problem. The proposed methodology, which imitates the human ability to process information, acquires knowledge from Information Granules rather then from numerical data. This method also introduces a Latent Semantic Indexing based feature extraction tool by using Singular Value Decomposition, to dramatically reduce the data dimensions. In addition, several data sets from the UCI Machine Learning Repository are employed to demonstrate the effectiveness of our method. Experimental results show that our method can significantly increase the ability of classifying imbalanced data.

1. Introduction

In recent years, we have seen an increase in research activities in the class imbalance problem. This increased interest resulted in two workshops being held, one by AAAI (American Association for Artificial Intelligence) in 2000 and another one by International Conference on Machine Learning (ICML) in 2003. *SIGKDD Explorations* also published one special issue in 2004. The problem is caused by imbalanced data, in which one class is represented by a large number of examples while the other is represented by only a few [4]. Imbalanced data will result in a significant bottleneck in the performance attainable by standard learning methods [20,27] which assume a balanced class distribution as shown in Fig. 1. It is regarded as one of the most relevant topics of future machine learning researches.

When learning from imbalanced data, traditional data mining methods tend to produce high predictive accuracy for the majority class but poor predictive accuracy for the minority class [39,45]. That is because traditional classifiers seek accurate performance over a full range of instances. They are not suitable to deal with imbalanced learning tasks [6,12,19,23] since they tend to classify all data into the majority class, which is usually the less important class. Fig. 2 illustrates this situation. If data mining approaches cannot classify minority examples such as medical diagnoses of an illness, or the abnormal products

[*] Corresponding author.
E-mail addresses: ittchen@mail.nctu.edu.tw (M.-C. Chen), lschen@cyut.edu.tw (L.-S. Chen), cchsu@cyut.edu.tw (C.-C. Hsu).

0020-0255/$ - see front matter © 2008 Elsevier Inc. All rights reserved.
doi:10.1016/j.ins.2008.03.018

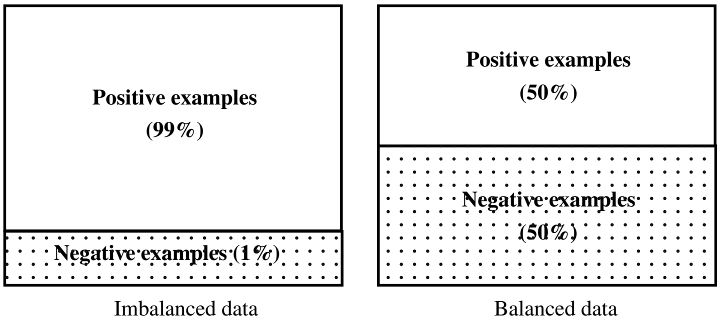

Fig. 1. Imbalanced and balanced data sets.

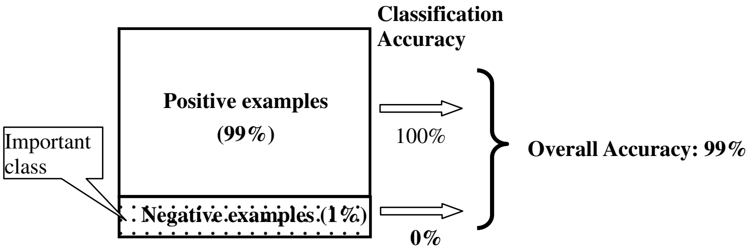

Fig. 2. The illustration of class imbalance problems.

of inspection data, the extracted knowledge becomes meaningless and useless. Recently, this problem has been recognized in a large number of real world domains, like medical diagnosis [37], inspection of finished products [38], identifying the cause of power distribution faults [39], surveillance of nosocomial infections [14], prediction of the localization sites of protein [3], speech recognition [25], credit assessment [20], and functional genomic applications [41].

To address the class imbalance problem, two major groups of techniques are proposed in the available literature. The first group involves five approaches: (1) *under-sampling*, a method in which the minority population is kept intact, while the majority population is under-sampled; (2) *over-sampling*, methods in which the minority examples are over-sampled so that the desired class distribution is obtained in the training set [6,11,19]; (3) *cluster based sampling*, methods in which the representative examples are randomly sampled from clusters [2]; (4) *moving the decision threshold*, methods in which the researcher tries to adapt the decision thresholds to impose bias on the minority class [11,21,24] and (5) *adjust costs matrices*, a method in which the prediction accuracy is improved by adjusting the cost (weight) for each class [15]. Besides, Liu et al. [27] also presented a weighted rough set method for this problem. However, all of these techniques have some disadvantages [2]. For instance, the computational load is increased and overtraining may occur due to replicated samples in the case of over-sampling. Under-sampling does not take into account all available training data which corresponds to loss of available information. Huang et al. [21] indicated that these supervised methods lack a rigorous and systematic treatment of the imbalanced data.

The second group is related to Granular Computing (GrC) models. These GrC models [37,38] which copy the human instinct of information processing can increase classification performance by improving the class imbalance situation. However, these models use the concept of sub-attributes to describe Information Granules (IGs) which are collections of objects arranged together based on their similarity, functional adjacency and indistinguishability [5,9,40,42]. When handling continuous data, the drawback of sub-attributes is that computational loads will increase dramatically due to the generation of a huge number of sub-attributes. Therefore, by introducing the Latent Semantic Indexing (LSI) based feature-extraction technique, this study proposes a novel GrC model called the “Information Granulation Based Data Mining Approach” to solve the class imbalance problem. In addition, for highly skewed data, we present a new IG construction strategy which only builds IGs from majority examples and keeps minority instances intact. Finally, the experimental results show the superiority of our method for classifying imbalanced data.

2. Granular computing

Humans have a remarkable capability to perform a wide variety of physical and mental tasks without any measurements/computations, such as playing computer game, driving, and cooking. Human beings use perceptions of direction, speed, time and other attributes of physical/mental objects, instead of numerical data. Basically speaking, reflecting the limited ability of human brains, perceptions are inaccurate. In more concrete terms, perceptions are granular. It means that the boundaries of