International Journal of Innovative Computing, Information and Control
Volume **6**, Number **11**, November **2010**

ICIC International ©2010 ISSN 1349-4198
pp. 4979–4994

DEVELOPING A NOVEL TWO-PHASE LEARNING SCHEME FOR THE CLASS IMBALANCE PROBLEM

Long-Sheng Chen[1], Chun-Chin Hsu[2] and Yu-Shan Chang[1]

[1]Department of Information Management
[2]Department of Industrial Engineering and Management
Chaoyang University of Technology
168 Jifong E. Rd., Wufong Township, Taichung County 41349, Taiwan
{ lschen; cchsu }@cyut.edu.tw

Received June 2009; revised November 2009

Abstract. *In classification problems, the class imbalance problem will cause a bias on the training of classifiers and result in a low predictive accuracy over the minority class examples. Lots of real-world data sets have skewed class distributions in which almost all examples belong to one class and only a few instances belong to others. Usually, these minority examples are a class of higher interest, such as rare diseases in medical diagnosis, defectives in failure monitoring, frauds in credit screening, and so on. In order to tackle the class imbalance problem, this study aimed to (1) find a robust classifier from different candidates including Decision Tree (DT), Logistic Regression (LR), Mahalanobis Distance (MD), and Support Vector Machines (SVM); (2) propose a new two-phase learning scheme, called the MD-SVM methodology. The experimental results indicated that our proposed MD-SVM has a better performance in identifying the minority class examples compared to the traditional techniques such as under-sampling, cost adjusting, and cluster based sampling.*

Keywords: Class imbalance problem, Classification, Mahalanobis distance, Data mining, Support vector machines

1. Introduction.

In classification problems, the class imbalance problem causes a bias in the training of classifiers resulting in lower sensitivity for detecting the minority class examples [6]. This problem is caused by imbalanced data in which almost all examples belong to one class and only a few instances belong to others. A classifier induced from an imbalanced data set has high classification accuracy for the majority class, but an unacceptable error rate for the minority class [6,9,17,25]. This is because traditional classifiers seek an accurate performance over a full range of instances [38] and thus, the examples in the minority class are treated as noise and are ignored completely by the classifier [6]. Lots of real-world data sets have skewed class distribution such as medical diagnosis, failure inspection/monitoring, credit screening, text classification and others [4-6,25,34]. Recently, researchers in the machine learning and data mining community have been paying increasing attention to this problem and how it affects the learning performance of some well-known classifiers [6,17,25,29,34].

To tackle the class imbalance problem, many approaches have been presented. Generally speaking, there are two groups of methods for solving the class imbalance problem. They are the algorithm/models oriented approaches and the re-sampling techniques. The purpose of the former is to propose new learning mechanism/strategies or to modify the existing methods. Here, researchers attempt to solve the class imbalance problems by presenting different methods such as one class learning [7,23], support vector machines