Borderline-SMOTE: A New Over-Sampling Method in Imbalanced Data Sets Learning

Hui Han[1], Wen-Yuan Wang[1], and Bing-Huan Mao[2]

[1] Department of Automation, Tsinghua University, Beijing 100084, P. R. China
hanh01@mails.tsinghua.edu.cn
wwy-dau@mail.tsinghua.edu.cn
[2] Department of Statistics, Central University of Finance and Economics, Beijing 100081, P. R. China
maobinghuan@yahoo.com

Abstract. In recent years, mining with imbalanced data sets receives more and more attentions in both theoretical and practical aspects. This paper introduces the importance of imbalanced data sets and their broad application domains in data mining, and then summarizes the evaluation metrics and the existing methods to evaluate and solve the imbalance problem. Synthetic minority over-sampling technique (SMOTE) is one of the over-sampling methods addressing this problem. Based on SMOTE method, this paper presents two new minority over-sampling methods, borderline-SMOTE1 and borderline-SMOTE2, in which only the minority examples near the borderline are over-sampled. For the minority class, experiments show that our approaches achieve better TP rate and F-value than SMOTE and random over-sampling methods.

1 Introduction

There may be two kinds of imbalances in a data set. One is between-class imbalance, in which case some classes have much more examples than others [1]. The other is within-class imbalance, in which case some subsets of one class have much fewer examples than other subsets of the same class [2]. By convention, in imbalanced data sets, we call the classes having more examples the majority classes and the ones having fewer examples the minority classes.

The problem of imbalance has got more and more emphasis in recent years. Imbalanced data sets exists in many real-world domains, such as spotting unreliable telecommunication customers [3], detection of oil spills in satellite radar images [4], learning word pronunciations [5], text classification [6], detection of fraudulent telephone calls [7], information retrieval and filtering tasks [8], and so on. In these domains, what we are really interested in is the minority class other than the majority class. Thus, we need a fairly high prediction for the minority class. However, the traditional data mining algorithms behaves undesirable in the instance of imbalanced data sets, as the distribution of the data sets is not taken into consideration when these algorithms are designed.

The structure of this paper is organized as follows. Section 2 gives a brief introduction to the recent developments in the domains of imbalanced data sets. Section 3

D.S. Huang, X.-P. Zhang, G.-B. Huang (Eds.): ICIC 2005, Part I, LNCS 3644, pp. 878 – 887, 2005.

describes our over-sampling methods on resolving the imbalanced problem. Section 4 presents the experiments and compares our methods with other over-sampling methods. Section 5 draws the conclusion.

2. The Recent Developments in Imbalanced Data Sets Learning

2.1 Evaluation Metrics in Imbalanced Domains

Most of the studies in imbalanced domains mainly concentrate on two-class problem as multi-class problem can be simplified to two-class problem. By convention, the class label of the minority class is positive, and the class label of the majority class is negative. Table 1 illustrates a confusion matrix of a two-class problem. The first column of the table is the actual class label of the examples, and the first row presents their predicted class label. TP and TN denote the number of positive and negative examples that are classified correctly, while FN and FP denote the number of misclassified positive and negative examples respectively.

Table 1. Confusion matrix for a two-class problem

	Predicted Positive	Predicted Negative
Positive	TP	FN
Negative	FP	TN

$$\text{Accuracy} = (\text{TP}+\text{TN})/(\text{TP}+\text{FN}+\text{FP}+\text{TN}) \tag{1}$$

$$\text{FP rate} = \text{FP}/(\text{TN}+\text{FP}) \tag{2}$$

$$\text{TP rate} = \text{Recall} = \text{TP}/(\text{TP}+\text{FN}) \tag{3}$$

$$\text{Precision} = \text{TP}/(\text{TP}+\text{FP}) \tag{4}$$

$$F-value = ((1+\beta^2)\cdot \text{Re}call \cdot \text{Pr}ecision)/(\beta^2 \cdot \text{Re}call + \text{Pr}ecision) \tag{5}$$

When used to evaluate the performance of a learner for imbalanced data sets, accuracy is generally apt to predict the majority class better and behaves poorly to the minority class. We can come to this conclusion from its definition (formula (1)): if the dataset is extremely imbalanced, even when the classifier classifies all the majority examples correctly and misclassifies all the minority examples, the accuracy of the learner is still high because there are much more majority examples than minority examples. Under the circumstance, accuracy can not reflect reliable prediction for the minority class. Thus, more reasonable evaluation metrics are needed.

ROC curve [9] is one of the popular metrics to evaluate the learners for imbalanced data sets. It is a two-dimensional graph in which TP rate is plotted on the y-axis and