45

Data Mining for Imbalanced Datasets: An Overview

Nitesh V. Chawla

Department of Computer Science and Engineering
University of Notre Dame
IN 46530, USA
nchawla@cse.nd.edu

Summary. A dataset is imbalanced if the classification categories are not approximately equally represented. Recent years brought increased interest in applying machine learning techniques to difficult "real-world" problems, many of which are characterized by imbalanced data. Additionally the distribution of the testing data may differ from that of the training data, and the true misclassification costs may be unknown at learning time. Predictive accuracy, a popular choice for evaluating performance of a classifier, might not be appropriate when the data is imbalanced and/or the costs of different errors vary markedly. In this Chapter, we discuss some of the sampling techniques used for balancing the datasets, and the performance measures more appropriate for mining imbalanced datasets.

Key words: imbalanced datasets, classification, sampling, ROC, cost-sensitive measures, precision and recall

45.1 Introduction

The issue with imbalance in the class distribution became more pronounced with the applications of the machine learning algorithms to the real world. These applications range from telecommunications management (Ezawa et al., 1996), bioinformatics (Radivojac et al., 2004), text classification (Lewis and Catlett, 1994, Dumais et al., 1998, Mladenić and Grobelnik, 1999, Cohen, 1995b), speech recognition (Liu et al., 2004), to detection of oil spills in satellite images (Kubat et al., 1998). The imbalance can be an artifact of class distribution and/or different costs of errors or examples. It has received attention from machine learning and Data Mining community in form of Workshops (Japkowicz, 2000b, Chawla et al., 2003a, Dietterich et al., 2003, Ferri et al., 2004) and Special Issues (Chawla et al., 2004a). The range of papers in these venues exhibited the pervasive and ubiquitous nature of the class imbalance issues faced by the Data Mining community. Sampling methodologies continue to be popular in the research work. However, the research continues to evolve with different applications, as each application provides a compelling problem. One focus of the initial workshops was primarily

O. Maimon, L. Rokach (eds.), *Data Mining and Knowledge Discovery Handbook*, 2nd ed.,
DOI 10.1007/978-0-387-09823-4_45,

the performance evaluation criteria for mining imbalanced datasets. The limitation of the accuracy as the performance measure was quickly established. ROC curves soon emerged as a popular choice (Ferri et al., 2004).

The compelling question, given the different class distributions is: *What is the correct distribution for a learning algorithm?* Weiss and Provost presented a detailed analysis on the effect of class distribution on classifier learning (Weiss and Provost, 2003). Our observations agree with their work that the natural distribution is often not the best distribution for learning a classifier (Chawla, 2003). Also, the imbalance in the data can be more characteristic of "sparseness" in feature space than the class imbalance. Various re-sampling strategies have been used such as random oversampling with replacement, random undersampling, focused oversampling, focused undersampling, oversampling with synthetic generation of new samples based on the known information, and combinations of the above techniques (Chawla et al., 2004b).

In addition to the issue of inter-class distribution, another important probem arising due to the sparsity in data is the distribution of data within each class (Japkowicz, 2001a). This problem was also linked to the issue of small disjuncts in the decision tree learning. Yet another, school of thought is a recognition based approach in the form of a one-class learner. The one-class learners provide an interesting alternative to the traditional discriminative approach, where in the classifier is learned on the target class alone (Japkowicz, 2001b, Juszczak and Duin, 2003, Raskutti and Kowalczyk, 2004, Tax, 2001).

In this chapter[1], we present a liberal overview of the problem of mining imbalanced datasets with particular focus on performance measures and sampling methodologies. We will present our novel oversampling technique, SMOTE, and its extension in the boosting procedure — SMOTEBoost.

45.2 Performance Measure

A classifier is, typically, evaluated by a confusion matrix as illustrated in Figure 45.1 (Chawla et al., 2002). The columns are the Predicted class and the rows are the Actual class. In the confusion matrix, TN is the number of negative examples correctly classified (True Negatives), FP is the number of negative examples incorrectly classified as positive (False Positives), FN is the number of positive examples incorrectly classified as negative (False Negatives) and TP is the number of positive examples correctly classified (True Positives). Predictive accuracy is defined as $Accuracy = (TP+TN)/(TP+FP+TN+FN)$.

However, predictive accuracy might not be appropriate when the data is imbalanced and/or the costs of different errors vary markedly. As an example, consider the classification of pixels in mammogram images as possibly cancerous (Woods et al., 1993). A typical mammography dataset might contain 98% normal pixels and 2% abnormal pixels. A simple default strategy of guessing the majority class would give a predictive accuracy of 98%. The nature of the application requires a fairly high rate of correct detection in the minority class and allows for a small error rate in the majority class in order to achieve this (Chawla et al., 2002). Simple predictive accuracy is clearly not appropriate in such situations.

[1] The chapter will utilize excerpts from our published work in various Journals and Conferences. Please see the references for the original publications.