Mining Data Streams with Skewed Distribution based on Ensemble Method

Yi Wang, College of Information Engineering, Northwest A&F University, Yangling, Shaanxi, China

ABSTRACT

In recent years, there have been some interesting studies on predictive modeling in data streams. However, most such studies assume relatively balanced and stable data streams but cannot handle well skewed (e.g., few positives but lots of negatives) and skewed distributions, which are typical in many data stream applications. In this paper, we propose an ensemble and cluster based sample method to deal with this situation. The study shows that this method has effective result on skewed data streams mining.

Keyword: Cluster-Sampling, Data Mining, Data Streams, Ensemble, Skewed Data Streams

1. INTRODUCTION

Many real applications, such as credit fraud, network intrusion detection, web click stream, generate continuously arriving data, as known as data streams (Babcock, Babu, Datar, Motwani, & Widom, 2002; Gao, Fan, Han, & Yu, 2007). In recent year, there are many useful methods to deal with the data streams. In these studies, researchers suppose that the data set are balanced. However, it is not practical in the real world, as the data set with imbalanced class distribution often occurs in classification and clustering scenarios when a portion of the classes possesses many more examples than others. In these cases, the positive instances are much less than negative instances. For example, the online credit card fraud rate of US is just 2% in 2006 (Gao, Fan, Han, & Yu, 2007). At the same time, the cost of misclassifying a credit card fraud will impose thousands of dollars loss on the bank. It’s quite necessary for us to study the skewed data streams.

In the problem of imbalanced data mining, the minority class, which we call positive instances, is far fewer than the majority class, we call negative instances. The researchers usually use the sample methods to tackle the problem (Drummond & Holte, 2003; van Hulse, Khoshgoftaar, & Napolitano, 2007). In our paper, we propose an cluster based sampling method to deal with this situation. Compared with basic under-sampling and over-sampling methods, the cluster based sample method can reserve
potentially useful majority-class examples, which the under-sampling method may discard. Meanwhile, the cluster based sample have no the overfitting problem, which often occurs in the oversample method. In this paper, we use the ensemble method with dynamic weighted majority to deal with the data streams. The study shows that this ensemble and cluster based sampling method has effective results on skewed data streams mining.

This paper is organized as following. Section 2 reviews the related work. Section 3 shows our ensemble framework for skewed data streams mining. The detailed experiment setting and results are shown in section 4, followed by our conclusion and future work in section 5.

2. RELATED WORK

There have been several strategies in handling imbalanced data sets. First approach is resizing training sets includes over-sampling minority class examples and under-sampling majority ones. Drummond and Holte (2003) test these two methods and find that the under-sampling outperforms over-sampling, because the over-sampling does not increase information, but it does lead to overfitting, which always make performance of the classifier poorly. Chawla, Hall, Bowyer, and Kegelmeyer (2002) use SMOTE method to balance the date sets, which is applicable when the data sets are highly imbalanced or there are very few examples of minority class. Yet this technique employs a lot of synthetic data for both minority and majority class cases, which is not applicable for data streams environment. The second approach emphasizes cost sensitive learning (Dietterich, Margineantu, Provost, & Turney, 2000; Elkan, 2001). In many real applications like credit fraud detection, medical diagnosis, making wrong decision is usually associated with very different costs. So, assigning different cost factors to false negatives and false positive will lead to better performance with respect to positive classes (Chawla, Japkowicz, & Kolecz, 2004). The ensemble approach consists of a set of individually trained classifiers whose predictions are combined to classify new instances. Hongyu and Herna (2004) use boosting and data generation method to improve performance of the skewed data set mining. Chen, Liaw, Breiman, (2004) use random forest to learn the skewed data sets. However, these ensemble methods only are suitable for the ordinary data sets.

In the recent years, many algorithms specifically tailored towards mining from data streams have been proposed. Using a sampling strategy based on Hoeffding bounds, the VFDT algorithm efficiently induces a decision tree in constant time (Domingos & Hulten, 2000). Meanwhile, it is generally believed that ensemble classifier could have better classification accuracy than a single classifier (Dietterich, 2000). The initial papers (Street & Kim, 2001; Wang, Fan, Yu, & Han, 2003) on classification data streams by ensemble methods use static majority voting (Street & Kim, 2001) and static weighted voting (Wang, Fan, Yu, & Han, 2003). However, to the best of our knowledge, there are not a lot of works so far on skewed data streams mining. Jing Gao (2007) proposed an ensemble framework to deal with this problem. The authors construct the ensemble classifiers on every batch. They use under-sampling and over-sampling methods to balance the distribution of both classes. Through preserving the positive examples in the past batch, this framework improves the classification of the positive class. But, there are several problems for us to discuss:

1. Because the framework preserves the positive examples in the past batches, as the data arrives, the size of positive examples become larger quickly, the skewed data streams may become the balanced data streams. At that time, the methods used for the skewed data streams may be useless.
2. In practical application, the data arrives quickly. Preserving the positive example may be unrealistic and impossible.
3. When concept changes, the positive examples preserved in the framework may give negative impact on the classification.
Constructing the Collaborative Supply Logistics Operation Mode in Assembly System under JIT Environment
[www.igi-global.com/article/constructing-collaborative-supply-logistics-operation/62290?camid=4v1a](www.igi-global.com/article/constructing-collaborative-supply-logistics-operation/62290?camid=4v1a)

Design and Implementation of the Embed Computer Based on CompactPCI Express Bus
[www.igi-global.com/article/design-implementation-embed-computer-based/59706?camid=4v1a](www.igi-global.com/article/design-implementation-embed-computer-based/59706?camid=4v1a)