Chapter 5

Unbalanced Sequential Data Classification Using Extreme Outlier Elimination and Sampling Techniques

T. Maruthi Padmaja
University of Hyderabad (UoH), India

Raju S. Bapi
University of Hyderabad (UoH), India

P. Radha Krishna
Infosys Lab, Infosys Limited, India

ABSTRACT

Predicting minority class sequence patterns from the noisy and unbalanced sequential datasets is a challenging task. To solve this problem, we proposed a new approach called extreme outlier elimination and hybrid sampling technique. We use k Reverse Nearest Neighbors (kRNNs) concept as a data cleaning method for eliminating extreme outliers in minority regions. Hybrid sampling technique, a combination of SMOTE to oversample the minority class sequences and random undersampling to undersample the majority class sequences is used for improving minority class prediction. This method was evaluated in terms of minority class precision, recall and f-measure on syntactically simulated, highly overlapped sequential dataset named Hill-Valley. We conducted the experiments with k-Nearest Neighbour classifier and compared the performance of our approach against simple hybrid sampling technique. Results indicate that our approach does not sacrifice one class in favor of the other, but produces high predictions for both fraud and non-fraud classes.

DOI: 10.4018/978-1-61350-056-9.ch005
INTRODUCTION

Unbalanced data classification is an important issue in today’s datamining community. There are several real world domains like intrusion detection; fraud detection and medical diagnosis (Visa & Ralescu, 2005) are unbalanced in nature. But some of these datasets like promoter recognition (Rani & Bapi, 2008), intrusion detection (Pradeep, Rao, Krishna, Bapi & Laha, 2005; Sanjay, Gulati & Pujari, 2004), and protein sequence prediction (Sikic, Tomic, & Vlahovicek, 2009; Zhao, Li, Chen, & Aihara, 2008) are sequential in nature, where each instance is the ordered list of discrete items. Unbalanced data classification problem is beatitude in those datasets when one class of data (majority class) severely outnumber the other class (minority class) of data.

We can solve imbalance problem that occur in sequence classification by using data mining techniques. If the imbalance problem is ignored and conventional classification methods are employed with the usual criterion of minimal overall error, then the model estimated will often ignore any contribution from the minority class samples. As such the model learned will only represent predominantly the majority class samples. These classification methods also assume that there is equal cost derived from all classes, which is not true in real world scenarios.

Consider intrusion detection system, (Pradeep, Rao, Krishna, Bapi & Laha 2005; Sanjay, Gulati, & Pujari, 2004) compared to non-intruder system call transactions, the occurrence of intruder transactions is infrequent. So it is extremely difficult to extract the intruder patterns in this scenario. In this work, we consider sequence classification as an unbalanced data classification problem where the majority samples outnumber the minority samples. Usually, the classification algorithms exhibit poor performance while dealing with unbalanced datasets and results are biased towards the majority class. Hence, an appropriate model is needed to classify unbalanced sequential data.

For these types of problems, we cannot rely upon the accuracy of the classifier because the cost associated with fraud sample being predicted as a non-fraud sample is very high. The performance measures that can be used here are cost based metrics, ROC analysis and minority class F-measure.

In this work we considered sequence classification as a binary classification problem and proposed a hybrid sampling approach called extreme outlier elimination with SMOTE and random undersampling. Here k Reverse Nearest Neighbors (kRNNs) concept is used as a data cleaning method for eliminating extreme outliers in minority regions before generating extra samples using SMOTE. Synthetic Minority Oversampling Technique (SMOTE) synthetically incorporates new samples in the distribution whereas random undersampling randomly deletes majority class samples from current distribution. Proposed approach is evaluated on a discrete sequential data set named Hill-Valley dataset. We identified optimal classifier based on its precision, recall and F-measure rates. Compared with other models constructed based on one-class classification techniques and other sampling techniques proposed approach yielded better performance.

The remainder of this chapter is organized as follows. In BACKGROUND section, we present the background of the proposed approach. The work related to proposed approach is discussed from two perspectives in Related Work section. Proposed approach and experimental setup is discussed in Solutions and recommendations section. Finally conclusions from the current work and future research directions are provided at the end of the chapter.

BACKGROUND

This section describes the background of the methods used for proposing extreme outlier elimination and Hybrid sampling approach.
Related Content

Robust Clustering with Distance and Density
[www.igi-global.com/article/robust-clustering-with-distance-and-density/181884?camid=4v1a](www.igi-global.com/article/robust-clustering-with-distance-and-density/181884?camid=4v1a)

Towards an Improved Ensemble Learning Model of Artificial Neural Networks: Lessons Learned on Using Randomized Numbers of Hidden Neurons
Fatai Anifowose, Jane Labadin and Abdulraheem Abdulraheem (2014). *Biologically-Inspired Techniques for Knowledge Discovery and Data Mining* (pp. 76-106).

A Semi-Automatic Annotation Method of Effect Clue Words for Chinese Patents Based on Co-Training
[www.igi-global.com/article/a-semi-automatic-annotation-method-of-effect-clue-words-for-chinese-patents-based-on-co-training/215003?camid=4v1a](www.igi-global.com/article/a-semi-automatic-annotation-method-of-effect-clue-words-for-chinese-patents-based-on-co-training/215003?camid=4v1a)

Organic Knowledge Management for Web-Based Customer Service
[www.igi-global.com/chapter/organic-knowledge-management-web-based/27910?camid=4v1a](www.igi-global.com/chapter/organic-knowledge-management-web-based/27910?camid=4v1a)