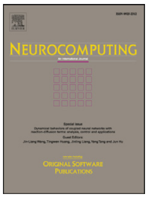
Neurocomputing 000 (2017) 1–14

Contents lists available at ScienceDirect

Neurocomputing

journal homepage: www.elsevier.com/locate/neucom



SOM-based partial labeling of imbalanced data stream

Elaheh Arabmakki*, Mehmed Kantardzic

Data Mining Lab, Department of Computer Engineering and Computer Science (CECS), Duthie Center for Engineering, University of Louisville, 222 Eastern Pkwy, Louisville, KY 40208, USA

ARTICLE INFO

Article history:
Received 14 July 2016
Revised 13 September 2016
Accepted 23 November 2016
Available online xxx

Keywords:
Data stream
Concept drift
Self-organizing map (SOM)
Support vector machines (SVM)
Partial labeling

ABSTRACT

Data streams are found in many large-scale systems such as security, finance, and internet. In many of the data streams, the class distribution is imbalanced, and hence most of the traditional classification modeler fails to produce high accuracy for the samples from minority class. In addition, data streams are changing and the model should be updated to maintain the classification performance. However, obtaining the true class labels of the data samples is not an easy task since labeling process is extremely time-consuming and very often class labels are not available immediately after classification. The goal of this research is to reduce the labeling for an imbalanced data stream, and to produce high classification performance when compared to fully labeling setting. In an imbalanced data stream, the challenging part is to find and label minority class samples. In this paper, we propose RLS-SOM (Reduced labeled Samples-Self Organizing Map) framework for classification of the imbalanced data stream in a non-stationary environment. RLS-SOM locates the minority class samples in the feature space using SOM. It maintains an ensemble of the classifiers and builds a new model when the changes occur, using only partial labeled samples. In RLS-SOM, the classification results are obtained from the ensemble, as well as each individual model in the ensemble. An individual model classification results are selected over ensemble results, if its performance is higher than the ensemble's performance. This comparison is performed to improve the performance as there may be one model in the ensemble that produces higher performance than the ensemble. Our experimental results demonstrate that RLS-SOM obtains higher performance when it is compared with several partially labeling techniques over benchmark data sets. In addition, the experimental results with other state of the art fully labeling methods such as UCB, SERA, SEA, and Learn++.CDS shows RLS-SOM maintains equivalent classification performance by using 10–30% labeling, on average.

1. Introduction

There are many real world classification problems which involve imbalanced data streams such as credit card fraud detection, rare disease diagnosis, and network intrusion detection [1]. In imbalanced data stream the number of samples that represent one class is much lower than the ones in the other class. The former is called minority class, and the latter is majority class [2]. For example in the case of credit card fraud detection, there are many legitimate transactions, compared to a few fraud ones. In rare disease detection, there are few patients that are diagnosed with a rare disease among large number of patients, or in network intrusion detection, there are abundant normal behaviors compared to few abnormal ones. In many of the applications, usually the correct classification of the minority class samples is of much importance [3].

In dynamically changing environments, the data distribution can change over time yielding the phenomenon called concept drift [1,4]. For instance, in the network intrusion detection, concept drift occurs due to the changes in the users' behavior, changes in the characteristics of legitimate users, or because of new creative adversary actions [5]. If we consider two consecutive time intervals t_0*, and* t_1, concept drift between time point t_0 and time point t_1 can be defined as follows:

$$\exists X : p_{t_0}(X,Y) \neq p_{t_1}(X,Y) \quad (1)$$

where p_{t0} denotes the joint probability distribution at time t_0 between the set of features X and the class Y. *P(X,Y)* is defined as follows:

$$p(X,Y) = p(Y|X).p(X) \quad (2)$$

where *p(X)* is the probability of the feature vector and *p(Y|X)* is the conditional probability of the class Y given feature vectors X. There are three ways in which concept drift may occur [3]:

* Corresponding author.
E-mail addresses: e0arab01@louisville.edu (E. Arabmakki), mmkant01@louisville.edu (M. Kantardzic).

http://dx.doi.org/10.1016/j.neucom.2016.11.088
0925-2312/© 2017 Elsevier B.V. All rights reserved.

- Gradual/Conditional Change: conditional probabilities *P(Y|X)* might change. This type of drift occurs since there is a change in the existing model boundary.
- Abrupt/Feature Change: *P(X)* might change. This is caused by samples appearing in the regions of data space that were previously unoccupied.
- Dual Change: both *P(X),* and *P(Y|X)* change.

In the wake of concept drift, a new classification model should be built to maintain the performance. There are two different approaches in dealing with a new model in analyzing a data stream:

1.1. Maintaining a single classification model

In many of the approaches for analyzing the data stream, a single model is maintained [6–9]. That is, when concept drift occurs, a new model is built and the old model is replaced with the new one; or maybe the old model is adjusted toward new model.

1.2. Maintaining an ensemble of the classification models

Unlike maintaining a single model, an ensemble of the classifiers is maintained in many of the research activities [10–12]. Each time a new model is built, it is added to an ensemble of the models to be used as a part of classification.

In analyzing a data stream, maintaining an ensemble of the classifiers and obtaining the classification from combined output of all the models in the ensemble shown to be more effective than maintaining only a single model [13]. However, in some cases such as recurring concept, one individual model in the ensemble may produce higher classification results than the ensemble. In such cases, the classification performance of that individual model may outperform the ensemble's performance.

Building a new model requires a set of new labeled samples. However, labels can be costly to obtain since a labor is involved in the labeling process. In data streams environment, labeling is needed often because of the concept drift. Therefore, designing frameworks which maintains the quality of the model using partial labeled samples are preferred.

In this paper, we propose RLS-SOM framework for classification of the imbalanced data streams. The framework works for both types of conditional and features changes, using Support Vector Machine (SVM), and Kohonen's Self-Organizing Map (SOM). RLS-SOM maintains an ensemble of the models and it improves the classification performance by using the information from both the ensemble and all the individual models in the ensemble. The main contributions of this paper are as follows:

- Proposing a classification framework for highly imbalanced data streams, and building a new model only when the concept drift occurs, by using partial labeling while maintaining the same quality as 100% labeling.
- Handling concept drift including both gradual drift and abrupt drift.
- Detection of the minority class samples for labeling in highly imbalanced data stream with labeling small number of samples.
- Evaluating of each individual model results in addition to the ensemble's results; and selection of one individual model's classification, if its classification performance outperforms the ensemble's performance.

The rest of the paper is organized as follows: Section 2 summarizes various research in the domain of imbalanced data stream with concept drift. In Section 3, we introduce our RLS-SOM framework for classification of the imbalanced data stream, using partial labeling. Data sets and the experimental set up are given in Section 4. Finally in Section 5, the experimental results and the comparison with other partially and fully labeling methodologies in this domain are presented followed by conclusions in Section 6.

2. Literature review

The previous research in the domain of imbalanced non-stationary data stream maybe divided mainly into two groups: a) some used full labeling in building the model, whereas b) some others used partial labeling. We briefly describe several researches in each domain.

2.1. Classification model based on full labeling

Recently, chunk-based learning for concept drifting data stream with class imbalance has received attention by community of researchers [1,7,8,11,12,14–16]. Learn^{++}.NSE (Learn^{++} for Non-Stationary Environment) was proposed in [16] to address the concept drift in the data stream. Authors used an ensemble that employs age-adjusted weighting mechanism and computes the final hypothesis using a weighted majority vote. Learn^{++}.NSE was not designed for the class imbalance, and is biased toward majority class samples in terms of existence of heavy class imbalance in the data stream. Authors later proposed an extension to Learn^{++}.NSE by adding Synthetic Minority Class Oversampling Techniques (SMOTE) [17] to address class imbalance problem. The approach is called Learn^{++}.CDS (Learn^{++} for Concept Drift with SMOTE) [15]. SMOTE has been recognized for its ability to learn from severe class imbalance. Each time a new data arrives in such dynamic environment, first SMOTE is used to make the distribution of the data balanced, by creating artificial minority class samples in the feature space. Learn^{++}.CDS then checks to see if the distribution of the data at time *t*, is differed from the one at time *t*−1. This is checked by evaluating existing ensemble on the most recent chunk of the data $D^{(t)}$. Then, the weights of the misclassified samples are increased and renormalized to create a penalty distribution. A new model is then will be built on the most recent data chunk $D^{(t)}$.

Uncorrelated Bagging (UCB) is proposed in [1] to address concept drifting imbalanced data stream. In this framework, to deal with class imbalance, the minority class samples are aggregated from the previous chunks .The majority class samples are undersampled from the current chunk and a bagging framework trains a model on the combined majority and minority class samples. In this approach, it may happen that the minority class samples become majority in the future due to the concept drift, and therefore these data samples are becoming irrelevant in the future. An extension of UCB proposed in [18] in which in addition to accumulating all minority class samples, the misclassified majority class samples from current model is also propagated. In this approach, the performance of the ensemble members was increased because the boundary between the classes was defined in a better way. Additionally, in this work to address the concept drift, a combination of information gain and Hellinger distance was used to determine the similarity of the current chunk to the other chunks of data. Thus each ensemble member's probability estimate is weighted by the similarity measure in order to obtain a more accurate prediction. The Hellinger distance is defined as follows:

$$HD(X,Y,f) = \sqrt{\sum_{v\in f}\left(\sqrt{\frac{|X_{f=v}|}{|X|}} - \sqrt{\frac{|Y_{f=v}|}{|Y|}}\right)^2} \tag{3}$$

where *X* and *Y* are two chunks of data for a given feature *f*. The information gain for a chunk *X* is defined as the decrease in entropy *H* of a class *c* conditioned upon a particular feature *f*.

$$IG(X,f) = H(X_c) - H(X_c \big| X_f) \tag{4}$$

Please cite this article as: E. Arabmakki, M. Kantardzic, SOM-based partial labeling of imbalanced data stream, Neurocomputing (2017), http://dx.doi.org/10.1016/j.neucom.2016.11.088