Advanced Science and Technology Letters
Vol.63 (NGCIT 2014), pp.192-195
http://dx.doi.org/10.14257/astl.2014.63.42

Modeling for Classifying Data Streams with Concept Drift

Joung Woo Ryu[1], Jin-Hee Song[2]

[1] Technical Research Center, SafeTia Lt.Co., Republic of Korea,
[2] School of IT Convergence Engineering, Shinhan University, Rebublic of Korea
[1]ryu0914@gmail.com, [2]jhsong@shinhan.ac.kr,

Abstract. We propose a novel methodology of maintaining a classification model on streaming data, such as sensor data or log data. Our approach uses an ensemble for classifying data streams which consists of a set of classifiers. New classifiers of the ensemble on streaming data will be generated dynamically according to the estimated distribution of streaming data instead of periodically building them. Also our approach is able to deal with changes of class distribution as well as changes of data distribution. We compared the results of our approach and of the previous approach which can only deal with changes of data distribution. In experiments with 10 benchmark data sets, our approach produced an average of 3.61% higher classification accuracy.

Keywords: modeling classification, streaming data, ensemble, concept drift

1 Introduction

Many companies which deal with customer's preferences or life styles provide reliable services using a prediction/classification model management method. Those models have to predict/classify something from streaming data in real time. A major characteristic of streaming data is to be changed in data distribution according to data generation status. Usually prediction/classification models are periodically updated, and those methods have the following disadvantages [1]. First, if the accuracy of a classifier is high, the classifier does not need to be updated. However, even in this case, the classifier is updated because of the update period. Second, if the current time doesn't reaches a period setting value, the classifier would not be updated even though the distribution of streaming data is changed within the update period. Third, human experts should label all samples, and then evaluate or refine the current classifier using them. In a real-world application, it is very impractical process that a human expert gives the classifier feedback on its decision for every single sample.

We propose an efficient ensemble-based modeling approach for classifying data streams. Our approach is able to dynamically generate new classifiers for an ensemble on streaming data. It decides if streaming samples should be selected for building new classifiers not according to a time interval, but according to a change in the distribution of streaming data. In addition, our approach can handle concept drift in an online process.

ISSN: 2287-1233 ASTL

2 Changes in Distribution of Streaming Data

A data stream is a continuous and infinite sequence of data, which makes either storing or scanning all the historical data nearly impossible [2]. Moreover, streaming data often evolve considerably over time [2]. The change in streaming data distribution is referred to as concept drift [3]. Types of the concept drift are categorized into (a) *the change of data distribution* or (b) *the change of class distribution* according to change in streaming data distribution, as shown in Fig. 1.

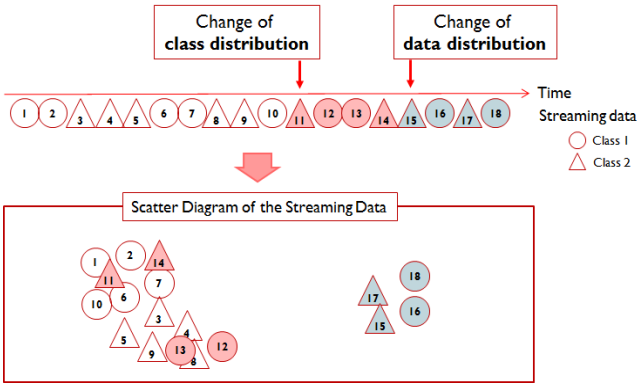

Fig. 1 Categorization of concept drifts according to change in streaming data distribution

3 Ensemble-based Modeling for Classifying Data Stream

We propose a more flexible approach which does not build a new classifier periodically in a fixed interval of time. Our ensemble approach decides dynamically when to build a new classifier and which samples should be used as a training data set according to changes in the distribution of streaming data.

Our approach adopts the concept of training data areas in order to estimate the current streaming data distribution. A training data area is defined the mean vector (*center*) and the standard deviation of a training data set from which a classifier is built. If the coming streaming data do not belong to any training data areas and the streaming data are near to each other, we can assume that a change in the current distribution of streaming data occurs. Using such an assumption, a methodology of maintaining the performance of an ensemble classifier in streaming data was proposed and its validity was showed by experiments with ten benchmark data sets [4]. However, the previous methodology proposed by [4] suffers from changes of class distribution occurring within training data areas. We improved the accuracy of the previous methodology by overcoming its disadvantage.

We assume that a classifier has a class distribution of its training data. Therefore, our approach compares the accuracy of a classifier on training data with its accuracy on test data. If the difference between the two accuracies is statistically significant, we

believe that a change of class distribution occurs. The test data set consists of data which randomly selected from streaming data belonging to the same training data area, and the selected streaming data will be labeled by human experts.

The size of a test data set is defined as the size of a training data set. If the size of a training data set is small, we decide whether the difference between the accuracies of the training data and the test data is statistically significant or not using the Fisher's Exact Test. The chi-square test is also used when the size of a training data set is large. In our experiments, the Fisher's Exact Test is used because the size of a training data set was defined as 300.

Our approach generates a meta-ensemble, as shown in Fig.2. An ensemble of a meta-ensemble is generated from a training data area. In other words, the ensembles of a meta-ensemble will be generated whenever a change of data distribution occurs. The classifiers of an ensemble will be built whenever a change of class distribution within a training data area is decided. In the meta-ensemble represented in Fig.2, we can know that once change of data distribution and changes of class distribution of three times occurred in streaming data so far.

When the meta-ensemble classifies a coming streaming datum, if the datum belongs to a training data area, it is classified by the only ensemble corresponding to the training data area. The final output value of an ensemble is decided by the simple majority voting method. If the datum does not belong to any training data areas, the meta-ensemble classifies the datum using all ensembles. The final output value of the meta-ensemble is decided by the weighted majority voting method. The weight values of each ensemble are calculated by the membership function used in Fuzzy C-means.

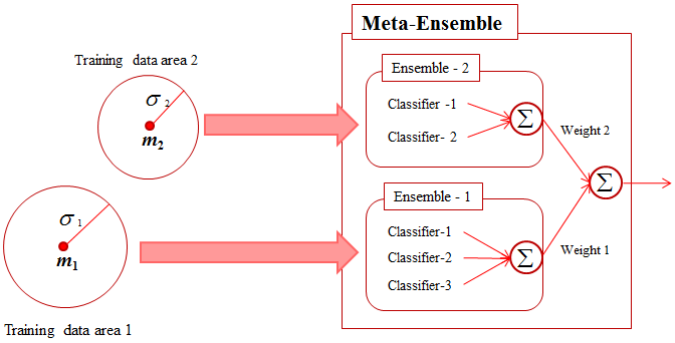

Fig. 2 Meta-Ensemble for classifying data streams

4 Experiments

We chose six real data sets from the UCI data repository and three streaming data sets from Wikipedia under the keyword "concept drift", as shown in Table 1.

We divided each data set into an initial training data set and a streaming data set. The initial training data set was used for building an initial ensemble which consists

of one classifier, the streaming data set was used as a test data set for evaluating ensemble approaches. Each value in Table 1 is presented with "the weighted sum of F-measures for all classes (WSF)", which is an appropriate measure for an ensemble accuracy applied for streaming data in a multi-classification problem with skewed class distribution [4]. Our ensemble's WSF value (0,621) is 3.61% higher than the WSF average for the previous ensemble [4].

Table 1. Comparison with WSF of our ensemble and the previsous ensemble[4]

Data set	Our ensemble	Previous ensemble [4]
Landsat Satellite	0.591	0.501
Mushroom	0.898	0.887
Nursery	0.515	0.485
MAGIC	0.742	0.774
Adult	0.532	0.510
Covertype	0.386	0.328
EM	0.670	0.643
PAKDD	0.275	0.290
KDDCup	0.982	0.978
AVERAGE	0.621	0.599

5 Conclusions

This paper presents a new modeling approach for classifying data streams according to changes in streaming data distribution. The main characteristic of the proposed ensemble-based approach is to be able to deal with the change of class distribution as well as the change of data distribution. We are going to apply the proposed approach to activity recognition using smartphone's accelerometer.

References

1. Haixun Wang, Wei Fan, Philip S. Yu: Mining concept-drifting data streams using ensemble classifiers. In: Proceeding of the 9th ACM SIGKDD international conference on knowledge discovery and data mining, pp.228-235, ACM, New York (2003)
2. Ioannis Katakis, Grigorios Tsoumakas, Ioannis Vlahavas: Tracking recurring contexts using ensemble classifiers: an application to email filtering. Knowledge and Information Systems, Vol.22, No. 3, 371-391 (2010)
3. Indrė Žliobaitė: Learning under concept drift: an overview. Technical report, Vilnius University, Faculty of Mathematics and Informatics 1-36 (2009)
4. Joung Woo Ryu, Mehmed M. Kantardzic, Myung-Won Kim, Efficiently Maintaining the Performance of an Ensemble Classifier in Streaming Data. In: Geuk Lee, Daniel Howard, Jeong Jin Kang (eds.) Korea-Daejeon 2012. LNCS, vol. 7425, pp. 533-540. Springer, Heidelberg(2012)

