INTERNATIONAL JOURNAL OF RESEARCH IN COMPUTER APPLICATIONS AND ROBOTICS
ISSN 2320-7345

EFFICIENT ADAPTIVE PREPROCESSING WITH DIMENSIONALITY REDUCTION FOR STREAMING DATA

Saranya Vani.M[1], Dr. S. Uma[2], Sherin. A[3]

[1] PG Scholar, PG CSE Department, Hindusthan Institute of Technology, Coimbatore, Tamil Nadu, India
[2]Head of the Department, PG CSE Department, Hindusthan Institute of Technology, Coimbatore, Tamil Nadu, India
[3]PG Scholar, PG CSE Department, Hindusthan Institute of Technology, Coimbatore, Tamil Nadu, India

Abstract

A lot of research efforts has been dedicated towards making learning models adapt to changing environments over time as streaming data is subjected to concept drift very often. These adaptive learning systems often need to analyze datasets with huge dimensions as and when they adapt their model. Performing the classification tasks on huge datasets leads to compromising the performance of the learning model. In order to overcome such difficulty, the Principal component analysis, a dimensionality reduction technique is incorporated with the adaptive learning system. PCA helps to reduce the dimensions of data and to improve the accuracy of classification for the learning model. Another major issue faced by adaptive learning models is the problem of feature evolution detection. The existing drift detection techniques for data streams either do not address the feature-evolution problem or suffer from high false alarm rate and false detection rates in many scenarios. Hence a feature evolution detection technique with adaptive threshold is incorporated with the adaptive learning systems to identify the evolution of novel features more precisely. Thus the proposed system focuses on improving the accuracy of the adaptive learning model by reducing the dimensions of data and detecting novel features more precisely.

Keywords: Adaptive preprocessing, Dimensionality reduction, Feature evolution detection

INTRODUCTION

In real time applications as data is evolving over time learning models need to have opportunities to update or retrain them. If adaptive learning is not possible, then the classification task could be less accurate or at times incorrect. Most of the supervised learning approaches consider that data is already preprocessed and not much effort is spent on adapting the preprocessing along with adapting the learning model [8]. Several adaptive learning systems have been proposed to address this problem. Some of the systems adapt both the preprocessor module and the classification module simultaneously or they adapt the modules individually according to the requirements. Research works shows that both the approach seems to have some limitations as the learning

model introduces additional computation costs when it adapts to data distribution changes. To overcome the limitation, a feedback mechanism is introduced between the pre-processor and the classifier module in [2].

The data dealt with adaptive data mining systems are often distributed in high dimensional spaces. People working with them regularly confront the problem of dimensionality reduction, which is a procedure of finding intrinsic low dimensional structures hidden in the high dimensional observations [3]. Most of the known dimensionality reduction methods are examples of one of the two classes of methods such as feature selection and feature extraction. Many literature works exists for the dimensionality reduction techniques such as Multidimensional Scaling (MDS), ISOMAP, Local Linear Embedding (LLE) and many more. Principal component analysis is a traditional approach for reducing the dimensions of data.

Novel feature evolution detection is a major issue in the evolving data streams. The existing adaptive systems suffer from high false alarm rate of differentiating between novel feature evolution and concept drift [7] . Feature evolution is an intrinsic characteristic of streaming data [9]. Consider the text data stream, such as that occurring in a social network such as Twitter. In this case, new topics (features) may frequently emerge in the underlying stream of text messages.

In this paper we propose a superior adaptive preprocessing system which claims two major contributions. First the adaptive preprocessing system is combined with the dimensionality reduction technique PCA. Second feature evolution detection algorithm with adaptive threshold is added to identify the novel features more precisely. We apply our proposed system to the weather dataset and prove that our framework improves the accuracy of the classification task of the adaptive preprocessing system.

The remainder of the paper is organized as follows. Section 2 describes the proposed framework for the adaptive preprocessing system with dimensionality reduction. Section 3 shows the experimental results. Section 4 concludes the proposed work.

FRAMEWORK FOR ADAPTIVE PREPROCESSING WIH DIMENSIONALITY REDUCTION

The motivation of adaptive preprocessing is to decouple the adaptively of preprocessing and classification by introducing a feedback mechanism between the two modules. Deciding upon when to adapt preprocessing and when to adapt classification is done with a selection strategy. The selection strategy is takes into consideration the following four scenarios to decide upon decoupling adaptively of the learning model. In scenario S0 the data is stationary and there is no need to adapt. Hence the newly arrived data adds to the existing model. In scenario S1, there is no need to adapt the preprocessor, but the predictor needs to be adapted. Such a situation may occur when the input data does not change, but the relation between the input and the target variables changes. In Scenario S2, there is no need to adapt the predictor, but the preprocessor needs to be adapted. Such situation may occur, when the distribution of data does not change, but the noise on data changes. In Scenario S3, there is a need to adapt both the preprocessor and the predictor. This may occur when the input data distribution changes.

The main objective of the proposed system is to improve the accuracy of classification of the adaptive preprocessing system. To obtain the goal our framework combines the novel feature detection algorithm and the dimensionality reduction technique to the adaptive system. The Principal Component Analysis (PCA) is used to reduce the dimensions of data and Novel feature detection with adaptive threshold is used to identify new features evolving in data stream. The diagram shown below is the architecture for the proposed system.

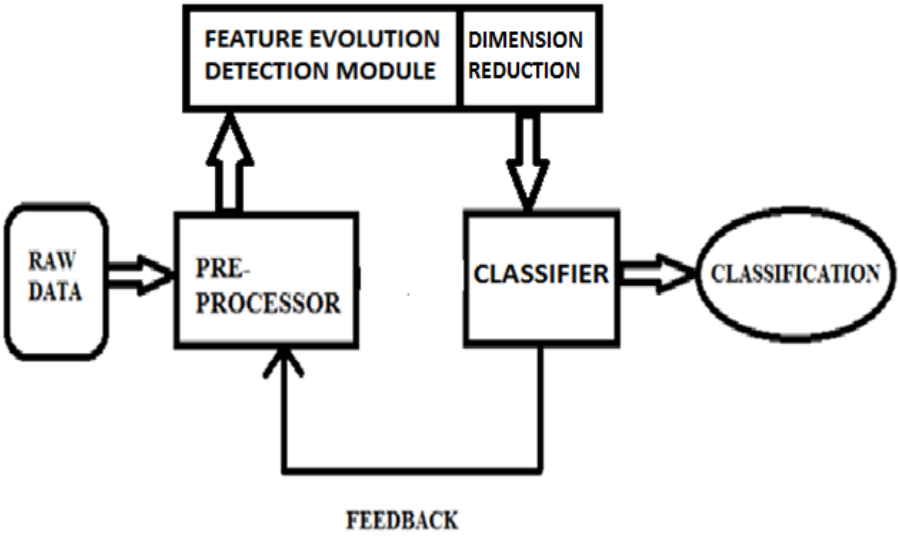

Fig 1 Architectural diagram of the Adaptive Preprocessing System with dimensionality reduction

Preprocessor

Real world data tends to be incomplete, noisy, and inconsistent and an important task when preprocessing the data is to fill in missing values smooth out noise and correct inconsistencies [6]. When dealing with missing data, there are several techniques that can be used. Choosing the right technique is a choice that depends on the problem domain and the goal for the data mining process. The techniques include ignoring the tuple with missing values, filling in the missing values manually, using a global constant to fill in the missing value, using the attribute mean to fill in the missing value, using the attribute mean for all samples belonging to the same class as the given tuple and using the most probable value to fill in the missing value.

Classification

Classification is a data mining technique used to predict group membership for data instances [5]. There are many traditional classification methods like decision tree induction, Bayesian networks, rule based classification, k-nearest neighbour classifier, support vector machines, case-based reasoning, genetic algorithm, fuzzy logic techniques, rough set approach and so on. The basic difference between the algorithms depends on whether they are lazy learners or eager learners. The decision tree classifiers, Bayesian classifier, support vector classifier are eager learners as they use training tuples to construct the data model whereas nearest neighbour classifiers are lazy learners as they wait until a test tuple arrives for classification to perform generalization [1].

Feature Evolution Detection

The feature detection algorithm identifies the outliers in the given dataset. A flexible decision boundary is set for outlier detection by allowing a slack space outside the decision boundary [4]. This space is controlled by a threshold, and the threshold is adapted continuously to reduce the risk of false alarms and missed novel classes. Each incoming instance in the data stream is first examined by a feature detection module to check whether it is a new feature. If it is not a new feature, then it is classified as an existing feature by the classifier. If it is an outlier, it is temporarily stored in a buffer. When there are enough instances in the buffer, the novel feature is detected and alarmed. We allow a slack space beyond the surface of each hypersphere. If any test instance falls within this slack space, then it is considered as existing class. This slack space is defined by a threshold, which is referred to as OUTTH. If this threshold is set too small, then the false alarm rate will go up, and vice versa. Therefore, we apply an adaptive technique to adjust the false alarm rate.

Dimensionality Reduction

The dimensionality reduction module reduces the dimensions of the given set of data. The traditional Principal Component Analysis is used to reduce the dimensions. The first principal component gives the data with maximum information about the data followed by the second and so on. Linear techniques perform dimensionality reduction by embedding the data into a subspace of lower dimensionality [10]. Although there exist various techniques to do so, PCA is by far the most popular (unsupervised) linear technique. Principal Components Analysis (PCA) constructs a low-dimensional representation of the data that describes as much of the variance in the data as possible. This is done by finding a linear basis of reduced dimensionality for the data, in which the amount of variance in the data is maximal. PCA has been successfully applied in a large number of domains such as face recognition, coin classification, and seismic series analysis.

EXPERIMENTAL RESULTS

The performance evaluation of the proposed system was done by comparing the same with the existing system. The experimentation was carried out by using the weather data from the archive www.cse.fau.edu/~xqzhu/Stream/sensor.arff. The weather data distribution or the goal of the weather classification is often subject to change over time (concept drift). Hence the weather dataset is effectively utilized to evaluate the proposed system.

The first step in the system is to import training data, train the pre-processor and the classifier. For experimental purpose we use mean as the pre-processor and decision tree as a classifier. Initially the pre-processor is trained for cleaning the data. Then the data is given to the decision tree algorithm to build the model for the training data. The decision tree classifier builds model dynamically for the given set of data.

The accuracy of the decision tree constructed is calculated by dividing the given same into training set and testing set. The training of the classifier is done with the training dataset and the testing set is used for evaluation. The accuracy of the model is evaluated using the cross validation of dataset and the testing errors is reported as mean square error. The mean square error is calculated by using the below formula

$$\frac{|p_1 - a_1| + \ldots + |p_n - a_n|}{n} \qquad \text{------ Eqn. (1)}$$

where p_i is the predicted target values and a_i the actual target values.

Training of data is followed by the detection of novel features in the incoming dataset. Novel features are identified by adjusting the adaptive threshold value. Similarly Principal component Analysis is used to reduce the dimensions of data. The first principal component yields the maximum information about the dataset followed by the second and so on. The weather dataset taken for experimental purpose consists of dataset of six attributes. The principal component analysis is applied to the dataset and the principal components (PC) are extracted. We take the first four PC's into consideration for our analysis. Thus the dataset with six dimensions is reduced to four dimensions.

After reducing the dimensions of data the model can be used to classify the test set. The online selection strategy is used for identifying the scenarios for adapting the pre-processor module and the classifier module. The strategy calculates the mean square error to classify the new data for all four strategies and selects the one with minimum error. The system is tested with different with different datasets and the accuracy of the system is compared with the existing one.

The below table shows the comparison between different datasets used for testing the accuracy of the existing system and the proposed system. The columns compare the accuracy of the existing and the proposed system.

Dataset	Cycle 1	Cycle 2	Cycle 3	Cycle 4
Preprocessor not adapted (Accuracy in percentage)	58 %	42 %	45%	50%
Preprocessor adapted (Accuracy in percentage)	58 %	59 %	59%	59%
Preprocessor adapted along with dimensionality reduction of data (Accuracy in percentage)	62%	60%	64%	59%

Table 1. Comparison of Accuracy between Existing and Proposed System

CONCLUSION

In this paper the need for adaptive preprocessing in evolving data and how it helps to improve the prediction accuracy is dealt with. An online selection strategy that handles adaptivity of preprocessing and adaptivity of predictor separately is implemented. The existing system does not have much opportunity to adapt to the environmental changes, whereas the proposed system identifies the scenarios under which it is beneficial to handle adaptivity of preprocessing and classification separately. The dimensions of data are further reduced to improve the accuracy of the classification task. A novel feature evolution detection technique is implemented to identify new evolving features in the evolving data stream. The proposed system is implemented and tested with the streaming weather dataset. The accuracy of the proposed system which adapts to the environmental changes with reduced dimensions of data is compared with the accuracy of the existing system which does not provide opportunities to adapt and it is proved that the proposed approach helps to improve the classification accuracy of the system as data evolves over time.

REFERENCES

[1] Data Mining Concepts and Techniques – Second edition, Jaiwei Han and MichelineKamber.
[2] Indre Ziobaite and Bogdan Gabrys, “Adaptive Preprocessing for Streaming Data”, IEEE transactions on knowledge and data Engineering, Feb 2014
[3] Kambhatla and T.K. Leen. Dimension reduction by local principal component analysis. Neural Computation, 9(7):1493–1516, 1997.
[4] Mohammad M. Masud, Charu C. Aggarwal, “Classification and Adaptive Novel Class Detection of Feature-Evolving Data Streams “IEEE transactions on knowledge and data Engineering, VOL. 25, NO. 7, July 2013
[5] Rutkowski, L.; Jaworski, M. ; Pietruczuk, L. ; Duda, P., “Decision Trees for Mining Data Streams Based on the Gaussian Approximation”- Knowledge and Data Engineering, IEEE Transactions on (Volume:26 , Issue: 1) Jan 2014
[6] Wang, H. Wang, X. Wu, W. Wang, and B. Shi, “A Low-Granularity Classifier for Data Streams with Concept Drifts and Biased Class Distribution,” IEEE Trans. Knowledge and Data Eng., vol. 19, no. 9, pp. 1202-1213, Sept. 2007.
[7] Wenerstrom and C. Giraud-Carrier, “Temporal Data Mining in Dynamic Feature Spaces,” Proc. Sixth Int’l Conf. Data Mining (ICDM), pp. 1141-1145, 2006.

[8] Yang, X. Wu, and X. Zhu, "Combining Proactive and Reactive Predictions for Data Streams," Proc. ACM SIGKDD 11th Int'l Conf. Knowledge Discovery in Data Mining, pp. 710-715, 2005.

[9] Zhang, X. Zhu, and L. Guo, "Mining Data Streams with Labelled and Unlabelled Training Examples," Proc. IEEE Ninth Int'l Conf. Data Mining (ICDM), pp. 627-636, 2009.

[10] [Teng, H. Li, X. Fu, W. Chen, and I.-F. Shen. Dimension reduction of microarray data based on local tangent space alignment. In Proceedings of the 4th IEEE International Conference on Cognitive Informatics, pages 154–159, 2005.

[11] Tsymbal. A, Pechenizkiy. M, Cunningham. P, and Puuronen. S, "Dynamic Integration of Classifiers for Handling Concept Drift," Information Fusion,Vol. 9, pp. 56-68, 2008.

[12] Xi-Zhao Wang; Ling-Cai Dong ; Jian-Hui Yan , "Maximum Ambiguity-Based Sample Selection in Fuzzy Decision Tree Induction", Knowledge and Data Engineering, IEEE Transactions on (Volume:24 , Issue: 8) - Aug2012

[13] Saul, K.Q. Weinberger, J.H. Ham, F. Sha, and D.D. Lee. Spectral methods for dimensionality reduction. In Semi supervised Learning, Cambridge, MA, USA, 2006. The MIT Press.

[14] Shawe-Taylor and N. Christianini. Kernel Methods for Pattern Analysis. Cambridge University Press, Cambridge, UK, 2004.

[15] Teng, H. Li, X. Fu, W. Chen, and I.-F. Shen. Dimension reduction of microarray data based on local tangent space alignment. In Proceedings of the 4th IEEE International Conference on Cognitive Informatics, pages 154–159, 2005.