Adaptive Pre-processing and Regression of Weather Data

Varsha Pullabhotla and K.P. Supreethi

Abstract With the evolution of data and increasing popularity of IoT (Internet of Things), stream data mining has gained immense popularity. Researchers and developers are trying to analyze data patterns obtained from various devices. Stream data have several characteristics, the most important being its huge volume and high velocity. Although, a lot of research is being conducted in order to develop more efficient stream data mining techniques, pre-processing of stream data is an area that is under-studied. Real time applications generate data which is rather noisy and contain missing values. Apart from this, there is the issue of data evolution, which is a concern when dealing with stream data. To deal with the evolution of data, the proposed solution offers a hybrid of preprocessing techniques which are adaptive in nature. As a result of the study, an adaptive preprocessing and learning approach is implemented. The case study with sensor weather data demonstrates the results and accuracy of the proposed solution.

Keywords Stream mining · Data evolution · Adaptive pre-processing

1 Introduction

In the present day scenario, there are many applications in our day to day lives ranging from social networks, health monitors, telecommunications, network monitoring tools, sensor devices (in manufacturing, industrial pumps etc.) and such which continually generate huge volume of data at high velocity. These data streams evolve over time. Thus, there is a need for adaptivity of predictive models

V. Pullabhotla (✉) · K.P. Supreethi
Computer Science and Engineering Department,
Jawaharlal Nehru Technological University Hyderabad, Hyderabad, Telangana, India
e-mail: varshapull28@gmail.com

K.P. Supreethi
e-mail: supreethi.pujari@jntuh.ac.in

H.S. Saini et al. (eds.), *Innovations in Computer Science and Engineering*,
Lecture Notes in Networks and Systems 8, DOI 10.1007/978-981-10-3818-1_2

to adapt to the evolution and change in environment of data streams. Recently, a lot of research and study is being carried out for such adaptive learning [1–3].

In real applications, pre-processing of data is a very important step of the data mining process as real data often comes from complex environments and can be noisy and redundant. In adaptive learning literature, the data pre-processing gets low priority in comparison to designing adaptive predictors. As data is continually changing, adapting only the predictor model is not enough to maintain the accuracy over time.

A good way to approach the above problem would be to tie the adaptivity of the preprocessor with the predictor. This can be accomplished in two ways. The first approach is to put aside a validation set, and use this validation set to optimize the pre-processing parameters and keep the pre-processing fixed in the model. The other approach would be to retrain the preprocessor afresh every time the learner is retrained. This approach requires the preprocessor to be synchronized with the learner.

In this paper, the aim is to present an implementation that can achieve adaptive pre-processing to get accurate output from adaptive learning. The pre-processing algorithm used is the "Multivariate Singular Spectrum Analysis". The learner algorithm used is the "K Nearest Neighbor" algorithm. These algorithms coupled with the Fixed window strategy [4] produce the adaptive pre-processing and learner framework.

The remainder of the paper is structured as follows: Sect. 2 presents the surveyed related work. Section 3 presents the proposed method for adaptivity. Section 4 shows the experimental results and performance evaluation. Finally, in Sect. 5, the conclusions drawn are presented.

2 Related Work

There has been a considerable amount of research and study conducted to address the issue of adaptive pre-processing along with adaptive learning. The issue of adaptive pre-processing while learning from a continuously evolving stream of data was raised in [4]. A framework that connects adaptive pre-processing to online learning scenarios was proposed. A prototype was developed to enable adaptive pre-processing and learning for stream data.

There has been the use of Genetic algorithm (GA) proposed by Wei Li to improve adaptive pre-processing to accomplish better results from adaptive learning [5].

Adaptive pre-processing of data streams has also been used with clustering algorithms. A pre-processing technique called equi-width cubes splits data space into a number of cubes, depending upon the data dimension and memory limit [6]. The new data which arrives is incorporated into one of the cubes. The algorithm computes a cluster center from previous chunk to create a new chunk. This new chunk is then sent to the clustering algorithm. This algorithm makes sure that the

data will not occupy all the available memory space and prevents loss of data due to the rate at which it arrives.

Adaptive pre-processing has been addressed in stationary online learning [7] for normalization of the input variables in neural networks. This was carried out so the input variables would fall into the range [−1, 1]. This proposed approach relates scaling of input features with scaling of the weights. However, the pre-processor is not adaptive.

3 Proposed Method

The framework of the proposed method is described in Fig. 1. The proposed Multivariate Singular Spectrum Analysis (MSSA) and K Nearest Neighbor approaches are applied to streaming weather data. Streaming weather data for the city of Hyderabad, India is used. The approach is carried out in two stages. In the first stage the MSSA algorithm which is used for pre-processing is trained with historical weather data for the city. The K-Nearest Neighbor algorithm is used for prediction. This model is trained using the output generated by the pre-processing algorithm. A stream of weather data is passed to this model, however, the results are not satisfactory.

The second stage involves applying the Fixed window strategy to the stream of weather data and retrain the preprocessor from scratch using the results obtained. The output obtained from the preprocessor results in the decomposition of the original time series into a stream of data without any noise. This output is passed to two K-nearest neighbor learner models: A model which is already trained using historical weather data and the other which is trained using the output obtained after retraining the pre-processor.

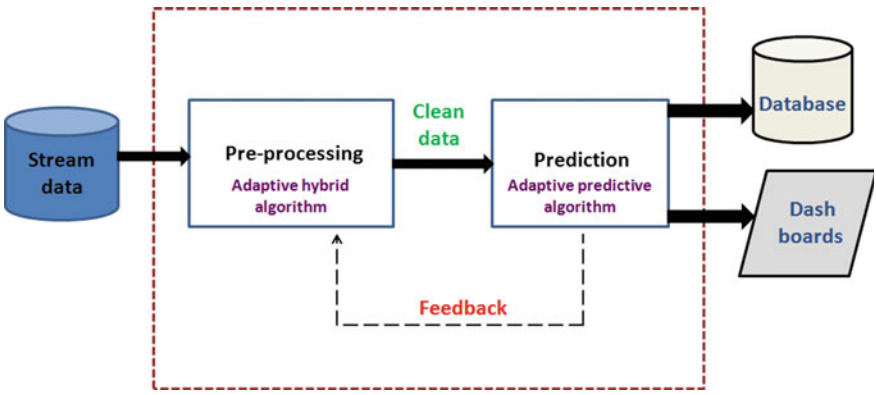

Fig. 1 Proposed system

4 Experimental Results

The proposed method is applied to streaming weather data for the city of Hyderabad, India. The performance measurement used to determine the accuracy of the prediction is the Root Mean squared error. This measure is meant to be used to understand how accurate the weather prediction for the next hour will be.

In this study, m = 3 and N = 10 are used, where N represents the sample size of the stream and m is the number of lags considered where the covariance is positive (this is determined using the autocorrelation function. We see a positive correlation at lag 3). The Root mean square error for the adaptive pre-processing and non-adaptive predictor is 0.29154. This error isn't considered too high and thus prediction is considerably accurate.

In the case of adaptive pre-processing and adaptive prediction, the RMSE is relatively low (0.014) and thus results in accurate prediction. However, this predictor is not trained by historical data as it is adaptive in nature. Thus this adaptive predictor has to be periodically re-trained for every 10 historical values to ensure that the predictor maintains its accuracy.

5 Conclusion and Future Scope

In this study, it has been demonstrated that the proposed approach, adaptive MSSA–KNN, could yield significantly higher prediction accuracy of weather data variables such as Temperature and Humidity than that of the non-adaptive KNN method. Adaptive MSSA–KNN results in a significant improvement prediction of weather data with RMSE of 0.014 and non-adaptive method results in RMSE of 0.29.

Future work would be focused on applying an incremental model instead of using a replacement model for adaptive pre-processing. Another area to work on would be to focus on passing multiple streams of weather data from different cities at once (Table 1).

Table 1 Prediction accuracies with adaptivity

Data	Adaptive pre-processing	Adaptive learning	RMSE
Stream weather data	No	No	4.008
Stream weather data	Yes	No	0.29154
Stream weather data	Yes	Yes	0.014

References

1. A. Bifet, G. Holmes, B. Pfahringer, R. Kirkby, and R. Gavalda.: New Ensemble Methods for Evolving Data Streams. In: Proc. 15th ACM SIGKDD Int'l Conf. Knowledge Discovery and Data Mining (KDD '09), pp. 139–148, 2009
2. E. Ikonomovska, J. Gama, and S. Dzeroski.: Learning Model Trees from Evolving Data Streams. In: Data Mining Knowledge Discovery, vol. 23, no. 1, pp. 128–168, 2011
3. P. Kadlec and B. Gabrys.: Architecture for Development of Adaptive on-Line Prediction Models. In: Memetic Computing, vol. 1, no. 4, pp. 241–269, 2009
4. Indrė Žliobaitė and Bogdan Gabrys.: Adaptive Pre-processing for Streaming Data. In: IEEE Transactions On Knowledge And Data Engineering, Vol. 26, No. 2, February 2014
5. Ketan Desale and Roshani Ade.: Preprocessing of Streaming Data using Genetic Algorithm. In: International Journal of Computer Applications (0975–8887) Volume 120–No.17, June 2015
6. Piotr Duda, Maciej Jaworski, and Lena Pietruczuk.: On Pre-processing Algorithms for Data Stream, L. Rutkowski et al. (Eds.): ICAISC 2012, Part II, LNCS 7268, pp. 56–63, 2012. Springer-Verlag Berlin Heidelberg 2012
7. H. Ruda.: Adaptive Preprocessing for on-Line Learning with Adaptive Resonance Theory (Art) Networks. In: Proc. IEEE Workshop Neural Networks for Signal Processing (NNSP), 1995

http://www.springer.com/978-981-10-3817-4

Innovations in Computer Science and Engineering
Proceedings of the Fourth ICICSE 2016
Saini, H.S.; Sayal, R.; Rawat, S.S. (Eds.)
2017, XVII, 378 p. 181 illus., Hardcover
ISBN: 978-981-10-3817-4