ISSN:0975-9646

Comparative Study of Pre-processing Techniques for Classifying Streaming Data

Ketan Desale[1], Roshani Ade[2]

[1,2] *Department of Computer Engineering*
Dr. D. Y. Patil School of Engineering & Technology
Savitribai Phule Pune University, Pune

***Abstract*— In today's world data is rapidly and continuously growing and is not constant in nature. There is a problem to deal with such kind of evolving data, because it is impractical to store and process this streaming data. Also, in real world application, the stream of data coming is typically noisy, has some missing values, repeated features, and thus very large time is wasted to process that data. The time complexity can reduce by selecting only useful features to build model for classification. The proposed system takes into consideration the issue of adaptive preprocessing for streaming data. Here Genetic algorithm (GA) is used as a search method while selecting the features which will further use in learning model. GA alongwith selective windowing strategy is the proposed system. The proposed system is applied to different stream datasets and, also compared with existing preprocessing technique PCA, is showing significant improvement in classification accuracy.**

***Keywords*— Genetic Algorithm, Streaming data, PCA, Preprocessing**

I. INTRODUCTION

The Data mining, also called Knowledge Discovery in Databases is the process of extracting useful information from the available data and has links with many fields like statistics, IR, machine learning and pattern recognition [1]. Data is generated and collected from many sources. Nowadays, we are also overwhelmed by data generated by computers and machines. Some of the examples are Internet routers, sensors, web servers, etc. This rapid generation of continuous streams of information has the limitation of storage, computation and communication capabilities in computing systems. To overcome these challenges, many models and techniques have been proposed over the past few years, known as adaptive learning. Adaptive learning model can be incremental or replacement. Incremental learning can be at the instance level, batch or ensemble level [2, 3]. Replacement can be full or partial.

Data mining is a complex topic and has links with multiple core fields such as statistics, IR, machine learning and pattern recognition. Data mining uses various tools such as classification, association rule mining, clustering. In real life scenarios, preprocessing is a very important factor of data mining process, because real data comes from a very complex environment and is often incomplete and redundant. In adaptive learning literature, the data preprocessing 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. Also, if we do not adapt preprocessing, the adaptive predictor may fail and in some cases give even worse results than nonadaptive predictors. The simple solution to automate preprocessing in adaptive learning can be to keep preprocessing tied with adaptive learners, which can be done in two cases. The first way is to make validations set at the start, optimize the preprocessing parameters on that validation set, and keep the preprocessing as it is for the rest of the model. The second way is to retrain all preprocessing from beginning every time the learner is retrained. This approach requires the synchronization of retraining of preprocessing and a predictor. One way to improve performance is to use a minimal number of features to define a model in a way that it can be used to accurately distinguish normal from anomalous behavior. Feature selection, also known as subset selection or variable selection, is a process usually used in machine learning, where a subset of the features of the available data is selected to use in a learning algorithm. Feature selection is an important task as it is computationally infeasible to use all available features for training the model. Wei Li described Genetic Algorithm based IDS with a methodology of applying genetic algorithm into network intrusion detection techniques.

In this paper, work is focused on improving adaptive preprocessing task so as to get the best output from adaptive learning. Feature selection using genetic algorithm is used as a preprocessing technique in an adaptive preprocessing model which gives relevant feature set.

II. RELATED WORK

The proposed system is designed by keeping goal to improve the performance of adaptive learning with the help of adaptive preprocessing. The majority of supervised learning methods assumes that the data comes already preprocessed or that preprocessing is an integral part of a learning algorithm. In real life applications, data which come from various sources is typically improper which contain missing values, redundant features. Thus more part of model development is utilized for data preprocessing. As data is evolving in nature, learning models also need to be able to adapt to changes dynamically.

A. *Adaptive Preprocessing*

The main goal of an adaptive system is to adapt to changes in data. Preprocessing does not operate individually as it is a part of adaptive systems. As the system is adaptive in nature, models that are used by the system change over time, with changes in data over time [4]. Zliobaite and Gabrys raised the issue of adaptive preprocessing in evolving data for the first time. Many supervised learning approaches that adapt to changes in data distribution over time have been developed for e.g. concept drift. A preprocessing component in adaptive prediction system has two main connections. First is the feedback, the preprocessor may need feedback from the predictor to adapt or retrain itself. Second is the mapping, the preprocessor produces a mapping that transforms the input data, which is then used by the predictor. The adaptivity of predictor may contaminate the feedback and by taking into consideration this the preprocessor decides when to adapt and whether adapt or not. This feedback is needed for updating of the preprocessor. At any given point in time, there may be a need to adapt the preprocessor or the predictor or both.

B. *Feature Selection*

It is the process of selecting a subset of the features of the available features to reduce dimensionality of the dataset [5, 6]. In FS redundant (duplicated valued) and irrelevant (contains no useful information) features are discarded. FS is an effective machine learning approach which further helps in building efficient classification system. With reduced feature subset, the time complexity is reduced with improved accuracy, of a classifier [7]. There are three standard approaches for feature selection: embedded, filter, and wrapper. In embedded approach FS occurs as a part of a data mining algorithm. Filter method selects features independent of the classifier used while in wrapper method features are selected specifically to classifier intended. Filter method uses any statistical way to while selecting features whereas wrapper uses a learning algorithm to find the best subset of features.

C. *Genetic Algorithm (GA)*

Genetic algorithms (GA) are an adaptive heuristic search method based on the idea of natural selection [8]. They are inspired by Darwins theory of evolution, survival of the fittest, which is one of the randomized search techniques. The algorithm begins with a set of individuals (chromosomes) called as population. Individual chromosome consists of a set of genes that could be bits, numbers or characters. Individuals are selected according to their fitness value for reproduction. Higher the fitness value more is the chances of an individual being selected [9]. Here mathematical intersection principle based innovative approach using genetic algorithm (GA) [10] is used as preprocessing technique.

Steps for GA:

1. Initialize the population P by randomly selecting individual form search space S.
2. Evaluate the fitness f(xi) for each individual in P
3. Repeat (until stopping condition satisfied)
   - Selection – according to the fitness value individuals are selected
   - Crossover – according to predetermined crossover probability, crossover the selected individuals
   - Mutation – according to mutation probability, newly generated in individuals are mutated Pnew
   - Update - P ← Pnew.
   - Evaluate – compute the fitness f(xi) of each individual in P
4. Return the most fitted individuals from P.

III. Proposed Approach

In this section, we use a simple prototype system for adaptive preprocessing. This comprises mechanism of selecting the most suitable approach dynamically using selective window strategy.

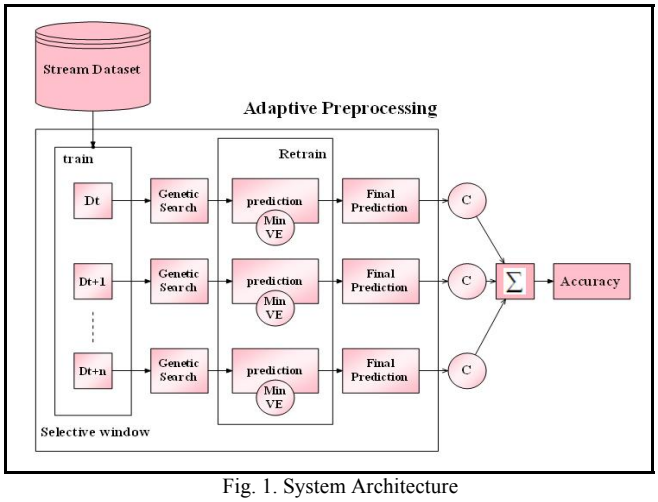

Fig. 1. System Architecture

Fig. 1 represents the proposed structure of the system for adaptive preprocessing. Here, system takes stream data as an input then it trains it after selecting window size dynamically by using selective window strategy. It contains two model preprocessing and prediction model. For preprocessing genetic algorithm is used as a search method while selecting relevant features. Then the strategy that has the minimum validation error is selected and the models retrained with this strategy. Finally summation of all the separate results is taken out for to get the final accuracy. This will return only required features set dynamically with respect to time, which will be considered for adaptive learning.

IV. Experimental Results

The experiment is carried on credit-g, NSL KDD train and test datasets [11] and self-constructed network dataset. The performance of proposed approach is compared with PCA preprocessing technique. The performance measurements used to compare classifier results are accuracy and a number of features selected.

TABLE I
ACCURACY WHEN MIN-50 AND MAX-150

Dataset	Without Preprocessing		With PCA		With Proposed Approach	
	Hoeffding	IBK	Hoeffding	IBK	Hoeffding	IBK
Credit-g	78.66	67.36	78.66	71.54	**81.76**	69.50
KDD Train	96.04	94.25	97.10	96.48	97.02	**99.05**
KDD Test	88.35	91.73	91.22	94.89	88.40	**95.54**
Own_Network	97.26	99.32	97.26	81.40	97.05	**99.55**

TABLE II
ACCURACY WHEN MIN-100 AND MAX-350

Dataset	Without Preprocessing		With PCA		With Proposed Approach	
	Hoeffding	IBK	Hoeffding	IBK	Hoeffding	IBK
Credit-g	78.66	67.36	78.66	71.54	74.05	**72.80**
KDD Train	96.04	94.25	97.10	96.48	96.04	**98.71**
KDD Test	88.35	91.73	91.22	94.89	88.75	**96.23**
Own_Network	97.26	99.32	97.26	81.40	96.99	**99.55**

TABLE III
ACCURACY WHEN MIN-150 AND MAX-250

Dataset	Without Preprocessing		With PCA		With Proposed Approach	
	Hoeffding	IBK	Hoeffding	IBK	Hoeffding	IBK
Credit-g	78.66	67.36	78.66	71.54	74.05	**72.80**
KDD Train	96.04	94.25	97.10	96.48	97.02	**98.55**
KDD Test	88.35	91.73	91.22	94.89	90.84	**95.54**
Own_Network	97.26	99.32	97.26	81.40	97.00	**99.45**

TABLE IV
FEATURES SELECTED

Dataset	Without Preprocessing		With PCA		With Proposed Approach	
	Hoeffding	IBK	Hoeffding	IBK	Hoeffding	IBK
Credit-g	21	21	21	21	4	4
KDD Train	42	42	42	42	3	3
KDD Test	42	42	42	42	4	4
Network	7	7	7	7	2	2

A. Accuracy

It means that how much our system is accurate enough to classify between correct and incorrect behaviour. The core decision factor in this system is window size. Here we calculated the results by varying the minimum and maximum window size. Three different sets of experiments are carried out according to window sizes. Table I to III are showing accuracy (in %) for four datasets by varying window size. From all results it is clearly showing the effect of proposed approach as preprocessing method for classifying streaming data.

B. Number of features selected

This performance measurement parameter shows us decrease in dimensionality in original dataset after applying proposed approach of feature selection. From Table IV we can see that proposed approach is selecting minimum number of features which further improves accuracy along with reduced time.

V. CONCLUSIONS

In this paper, mathematical intersection principle based innovative approach using genetic algorithm (GA) for feature selection is used for preprocessing streaming data along with selective window strategy. Proposed approach is compared with PCA technique on two different classifiers i.e. Hoeffding Tree and IBK. From the experimental results it can be concluded that the proposed method helps in selecting the minimum number of features which improves the classifier accuracy along with reduced time complexity. Amongst two classifiers IBK is giving good results in every case. Also proposed approach is giving good results for our created network dataset.

Future work would be focused on applying the proposed system to very high dimensional data which is also dynamic in nature.

REFERENCES

[1] Albert Bifet. 2009Adaptive Learning and Mining for Data Streams and Frequent Patterns. Doctoral Thesis, Universitat Politecnica de Catalunya

[2] Roshani Ade 2014 Instance based vs Batch based incremental learning approach for Students Classification. International Journal of Computer Application, Foundation of Computer Science, USA, vol. 106, no. 3

[3] Roshani Ade 2014 Classification of students by using an incremental ensemble of classifiers. 3rd IEEE International Conference On Reliability, Infocom Technologies and optimization, pp. 61-65, ICRITO- 8-10

[4] Indre Zliobaite, Bogdan Gabrys, "Adaptive Preprocessing for Streaming Data", IEEE Trans. Knowledge and Data Engg., vol. 26, no. 2, pp. 309- 321, Feb. 2014

[5] S Aksoy 2008 Feature Reduction and Selection Department of Computer Engineering, Bilkent University, 2008, CS 551

[6] B. Kavitha, S.Karthikeyan and B. Chitra 2010 Efficient Intrusion Detection with Reduced Dimension Using Data Mining classification Methods and Their Performance Comparison CCIS 70, pp. 96-101

[7] Mouaad KEZIH, Mahmoud TAIBI 2013 "Evaluation Effectiveness of Intrusion Detection System with Reduced Dimension Using Data Mining Classification Tools", 2nd International Conference on Systems and Computer Science (ICSCS) , August 26-27

[8] Wei Li 2004 "Using Genetic Algorithm for Network Intrusion Detection", Proceedings of the United States Department of Energy Cyber Security Grou, Training Conference, Vol. 8, pp. 24-27.

[9] Anup Goyal, Chetan Kumar, "GA-NIDS: A Genetic Algorithm based Network Intrusion Detection System".

[10] K. S. Desale, Rohani Ade, "Genetic Algorithm based Feature Selection Approach for Effective Intrusion Detection System", 2015 International Conference on Computer Communication and Informatics (ICCCI - 2015), Jan. 08 10, 2015, Coimbatore, INDIA

[11] NSL-KDD dataset for network-based intrusion detection systems available on http://iscx.info/NSL-KDD