Preprocessing of Streaming Data using Genetic Algorithm

Ketan Desale
ME Scholar
Dept. of Computer Engg.
Dr. D. Y. Patil school of Engg. & Tech.,
Savitribai Phule Pune University, Pune

Roshani Ade
Assistant Professor
Dept. of Computer Engg.
Dr. D. Y. Patil school of Engg. & Tech.,
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, as it is impractical to store and process this streaming data. Also, in real world application, the data which is coming is typically noisy, has some missing values, redundant features, and thus very large time is wasted to preprocess that data. The time complexity can reduce by selecting only useful features to build model for classification. The proposed system addresses 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. The proposed system is applied to different stream datasets and is showing significant increment in classification accuracy.

Keywords

Genetic Algorithm, streaming data, preprocessing

1. 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.

2. 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.

2.1 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.

2.2 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.

2.3 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].

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

3. PROPOSED APPROACH

The complete framework of the proposed approach is described in fig. 1. This paper describes a new approach of using genetic algorithm as evolutionary search method while feature selection. This proposed system is applied on two stream datasets and their results are analyzed.

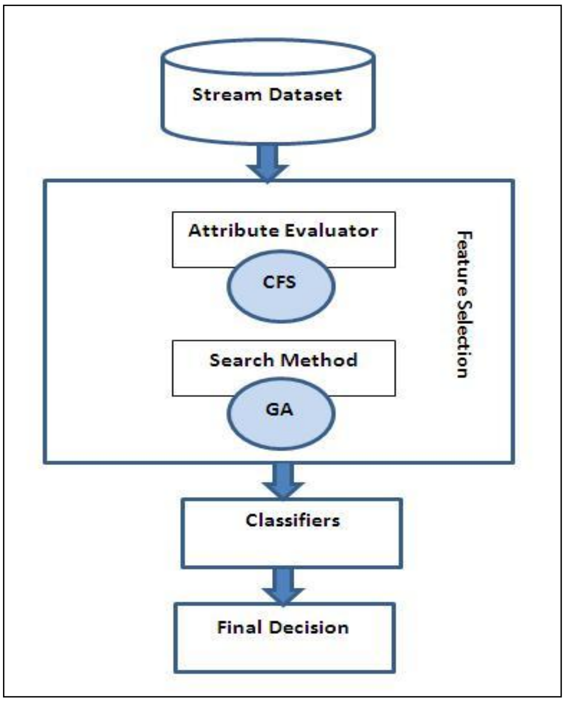

Fig 1: Proposed System

In this experiment, GA is run several times by changing the parameters, i.e. population size and maximum generations, while selecting features. For a specific population size experiment is carried out on three different generations. In all experiment crossover rate and mutation rate is kept constant. At the end intersection of the results of all experiment is taken out. This gives us only that features which are selected by each experiment. The population size indicates the chromosome number in one generation. Beyond the certain limit it is useless to increase the population size as it degrades the performance. Crossover rate is the probability of how many chromosomes will be used in reproduction [10]. The mutation rate is the probability of how often individuals will get muted.

4. EXPERIMENTAL RESULTS

The experiment is carried on both train and test NSL KDD datasets [12] and airline dataset [13]. The proposed approach is used with correlation based feature selection technique. The performance measurements used to compare classifier results are accuracy, time require building models and a number of features selected.

4.1 Accuracy

It means that how much our system is accurate enough to classify between normal and anomalous behavior. It is calculated as, Accuracy for KDD Train, KDD Test and airline datasets is as shown in fig 2, 3 & 4. The accuracy of the Naive Bayes classifier is notably increased for both the datasets after applying feature selection along with the proposed approach. At the same time j48 is showing performance degradation after feature selection.

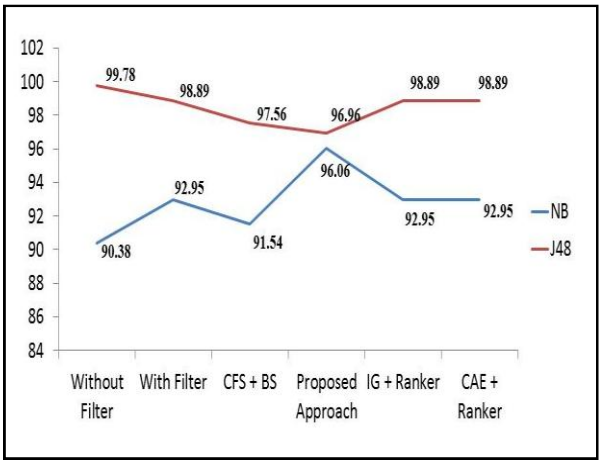

Fig 2: KDD Train accuracy (in %)

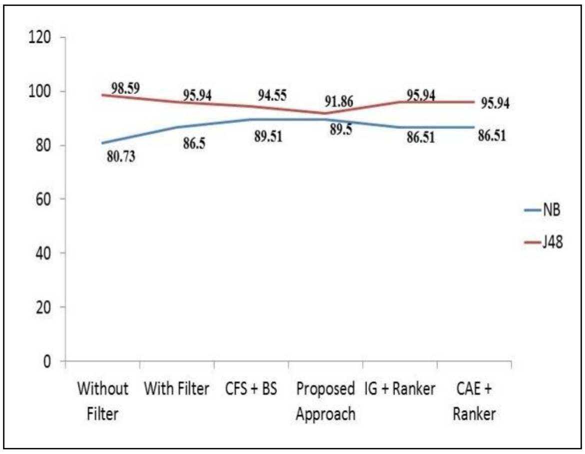

Fig 3: KDD Test accuracy (in %)

4.2 Time to Build

It is the time required to build the model. It is calculated in seconds. From fig 5 to 7 it is observed that CFS feature selection technique with Best First search takes less time but not improving accuracy. On the other hand CFS when used with a proposed approach takes slightly more time but with increased accuracy.

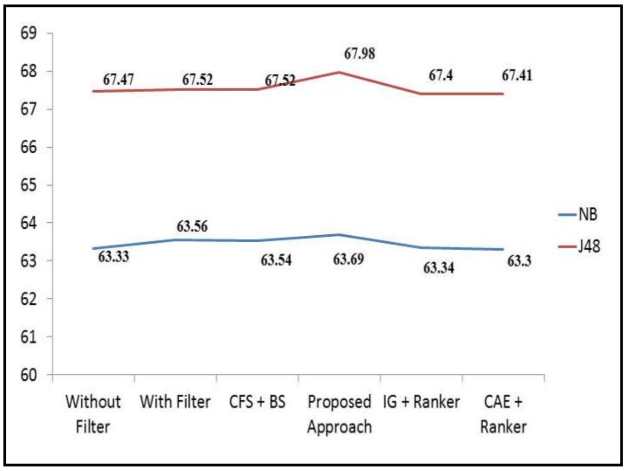

Fig 4: Airline accuracy (in %)

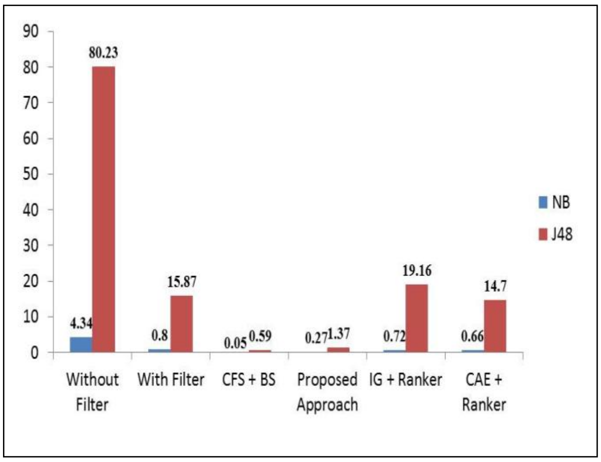

Fig 5: KDD Train Time to Build (in sec.)

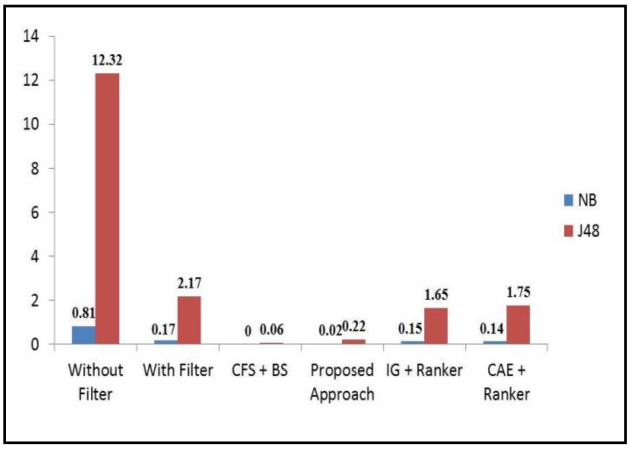

Fig 6: KDD Test Time to Build (in sec.)

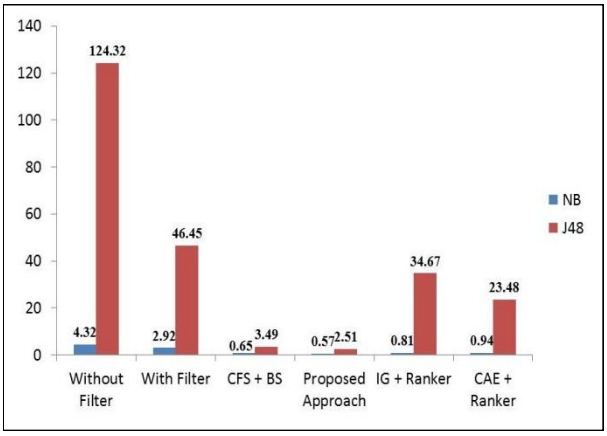

Fig 7: Airline Time to Build (in sec.)

4.3 Time to Build

This evaluation measure will help to analyze the dimensionality reduction factor. The original KDD dataset contains 41 attributes except class attribute. From the table below it is clearly observed that the proposed approach is selecting minimum number of features. KDD Train datasets have selected only 5 attributes out of 41 which mean about 88% dimensionality reduction is done. Similarly, for KDD Test dataset it has selected only 4 attributes from 41 attributes as shown in fig 8 which mean almost 90% dimensions get reduced. Similarly for airline dataset out of total 7 attributes except class attribute proposed approach selects only 3 attributes which leads to near about 43% dimensionality reduction.

Table 1. Selected Features

Dataset	**NB**				**J48**			
	CFS +BS	IG+ Ranker	CAE+ Ranker	Proposed	CFS+ BS	IG+ Ranker	CAE+ Ranker	Proposed
KDD Train	7	29	31	5	7	32	31	5
KDD Test	10	31	34	4	10	31	30	4
Airline	3	5	5	4	3	5	5	4

5. CONCLUSION AND FUTURE SCOPE

In this paper, mathematical intersection principle based innovative approach using genetic algorithm (GA) for feature selection is used for preprocessing streaming data. Feature selection is done using different feature selection (FS) techniques like CFS, IG and CAE and their effect on the performance of two commonly used classifiers, Naive Bayes and J48, is tested. From the experimental results it can be concluded that the proposed method helps in selecting the minimum number of features from both datasets i.e. NSL KDD & Airline data set which improves the Naïve Bayes classifier accuracy along with reduced time complexity.

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

6. 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] Bharat S. Dhak, Shrikant Lade, "An Evolutionary Approach to Intrusion Detection System using Genetic Algorithm", International Journal of Emerging Technology and Advanced Engineering, Volume 2, Issue 12, December 2012.

[11] 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

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

[13] Airlines data available on http://moa.cms.waikato.ac.nz/datasets/