An Efficient Approach to Enhance Classifier and Cluster Ensembles Using Genetic algorithms for Mining Drifting Data Streams

AnutoshPratap Singh
School Of Information Technology, R.G.P.V
Bhopal (M.P) India

JitendraAgrawal
School Of Information Technology, R.G.P.V
Bhopal (M.P) India

Varsha Sharma
School Of Information Technology, R.G.P.V
Bhopal (M.P) India

ABSTRACT

Mining data streams is concerned with extracting knowledge structures represented in models and patterns in high-speed streams of information. It raises new problems for the data mining community in terms of how to mine continuous high-speed data items that you can only have one look. The increasing focus of applications that generate and receive data streams stimulates the need for online data stream analysis tools. Recently, mining data streams with concept drifts has become an important and challenging task for a wide range of applications such as target marketing, network intrusion detection, credit card fraud protection, etc. Clustering and classification ensemble learning is a frequently used tool for building prediction models from data streams, due to its fundamental nature of managing large volumes of stream data. These both are the tools which help in improving the performance of mining systems. In order to improve the accuracy and error rate of traditional ensemble models, we propose a new ensemble model which combines both classifiers and clusters together and utilizes genetic algorithms for mining data streams. The main reason for using genetic algorithms along with clustering and classification is its high ability to solve optimization.

Keywords

Classifier and Cluster Ensembles Using Genetic algorithms,Genetic algorithms based Cluster and Classifier Ensembles for Mining Drifting Data Streams, Mining Data Streams.

1. INTRODUCTION

The rapid growth of continuous streams of data has challenged our storage, computation and communication capabilities in computing systems.Advances in networking and parallel computation have lead to the introduction of distributed and parallel data mining. The goal was how to extract knowledge from different subsets of a dataset and integrate these generated knowledge structures in order to gain a global model of the whole dataset. Client/server, mobile agent based and hybrid models have been proposed to address the communication overhead issue. Ensemble learning is a commonly used approach for mining Concept Drifting Data Streams [3].

Different variations of algorithms have been developed in order to increase the accuracy of the generated global model.

1.1 Classifier and Cluster Ensembles on data stream

Ensemble classifiers [1] on data streams provide a generic framework for handling massive volume data streams with concept drifting. The idea of ensemble classifiers is to partition continuous data streams into small data chunks, from which a number of base classifiers are built and combined together for prediction.

Whereas ensemble clusters [2] aims to combine multiple clusters together for prediction.Clustering ensemble is usually a two-staged algorithm. In the first stage, it stores the results of some independent runs of K-means or other clustering algorithms. In the secondstage, it uses a specific consensus function to find a final partition from the stored results.

2. GENETIC ALGORITHM

Genetic algorithms are the Artificial Intelligence technique. These algorithms are generally used to encode a potential solution to a specific problem. These algorithms are helpful in optimizing the result. The implementation of GA starts by with a population of typically random chromosomes. It means population is selected from huge amount of data on the random basis.working procedure of Genetic algorithm as shown. Genetic Algorithms are a family of computational models inspired by evolution. Fitness function is the key concept of this approach. All the calculations are done on the basis of this fitness function. If the chromosomes are able to survive in thesystem then it means that they satisfy the fitness function. There are basically three steps which are followed in the genetic process:

1. Selection
2. Crossover
3. Mutation

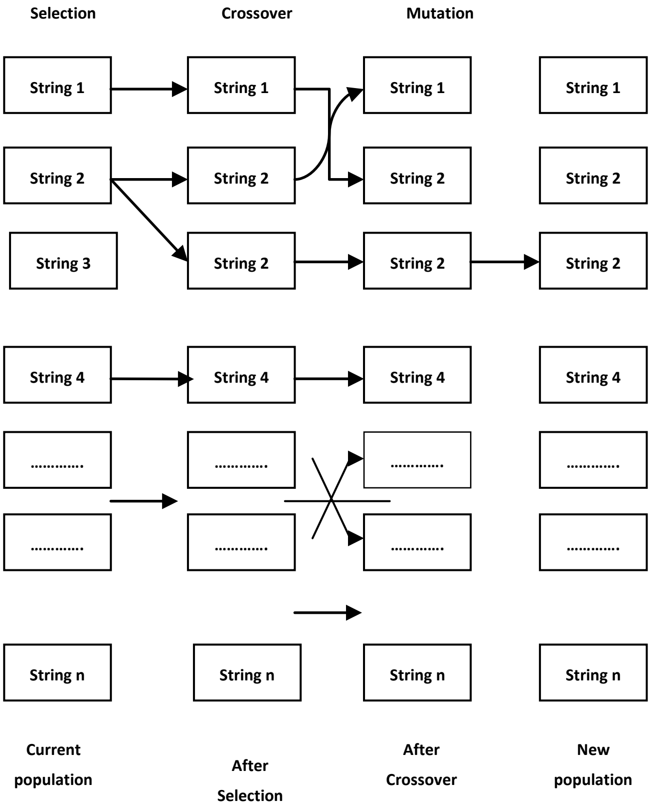


Figure 1: working procedure of Genetic algorithm

Working of Genetic algorithm is shown in figure 1. Genetic algorithm is also used in clustering approach. GA based clustering techniques [11] [12][13] are helpful in getting best optimized clusters.

3. RELATED WORK

There have been a lot of work done in the field of clustering and ensemble classification. But still there exist a problem in dealing with huge amount of data. Huge data volume and drifting concepts are not unknown to the data mining community. One of the goal of traditional data mining algorithms is to mine models from large databases. This complex goal is accomplished by some classification methods such as Sprint [4], Rainforest [5], BOAT [6], etc. but there is a problem that these algorithms require multiple scans of the training data which makes them inappropriate in the streaming data environment where the examples are coming in at a higher rate than they can be repeatedly analyzed. Incremental or online data mining methods [7] [6] are another option for mining data streams. Recently, a well-organized incremental decision tree algorithm called VFDT is introduced by Domingos et al [8]. This model works well with the streams, which are made up of discrete data. All the methods described above are producing a single model that represents the entire data stream. VFDT suffers in prediction accuracy in the presence of concept drifts.

In [1] author presents a label propagation method to deduce each cluster's class label by making full use of both class label information from classifiers, and internal structure information from clusters. To handle challenge, this paper presents a new weighting plan to weight all base models on the basis of their consistencies with the up-to-date base model. As a result, all classifiers and clusters can be combined together, through a weighted average mechanism, for prediction. Experiments on real-world data streams demonstrate that this process does better than simple classifier ensemble and cluster ensemble for stream data mining.

4. PROPOSED METHODOLOGY

In this paper we have used two previous Methods EC1 [9]and EC2 [10], for comparison purposes, which stand for Ensemble classifier. In EC1, we use Decision Tree as the basic learning algorithm. In EC2, we use logistic regression classifier, decision tree, and Naive Bayes as the basic algorithms. We are mainly dealing with Genetic algorithm in ensemble classification. Genetic algorithm is the branch of evolutionary algorithms, which help in improving the performance.

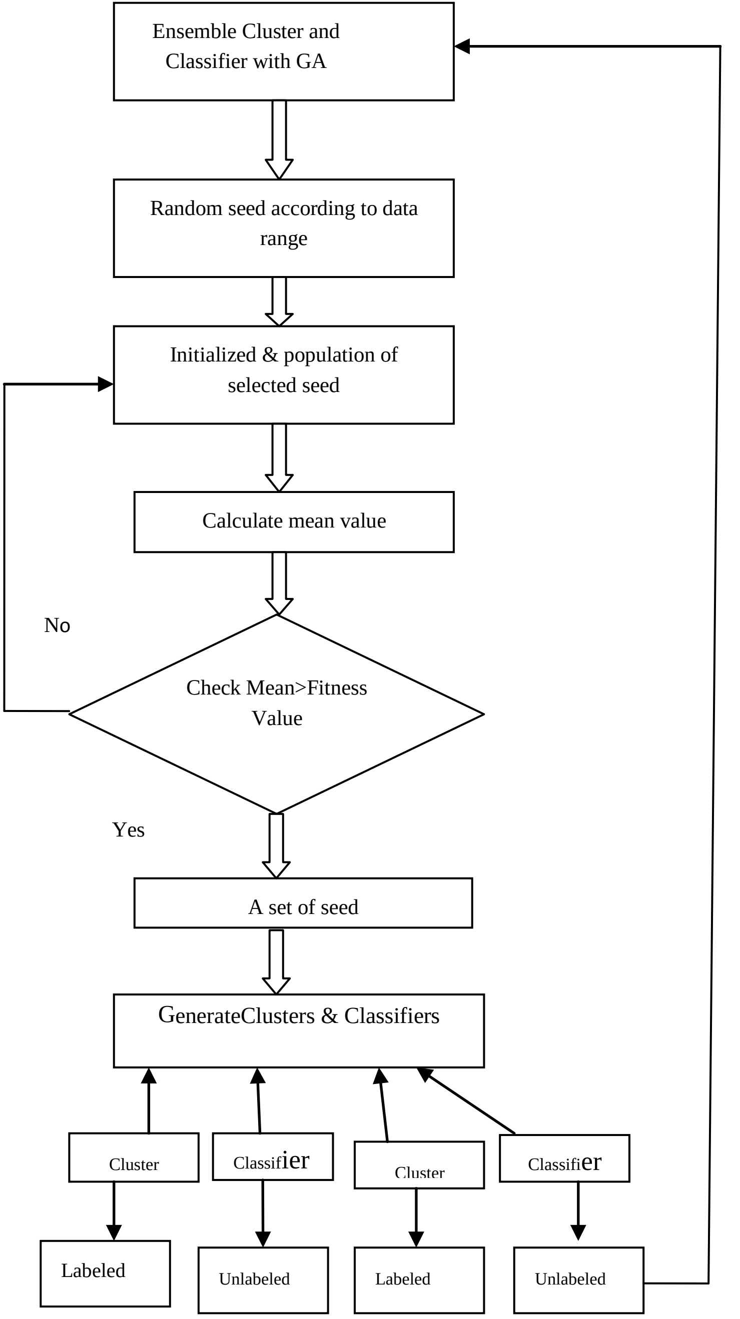

Ensemble Cluster and Classifier with GA
Random seed according to data range
Initialized & population of selected seed
Calculate mean value
No
Check Mean>Fitness Value
Yes
A set of seed
GenerateClusters & Classifiers
Cluster
Classifier
Cluster
Classifier
Labeled
Unlabeled
Labeled
Unlabeled
Data stream

Figure 2: Block diagram of proposed methodology

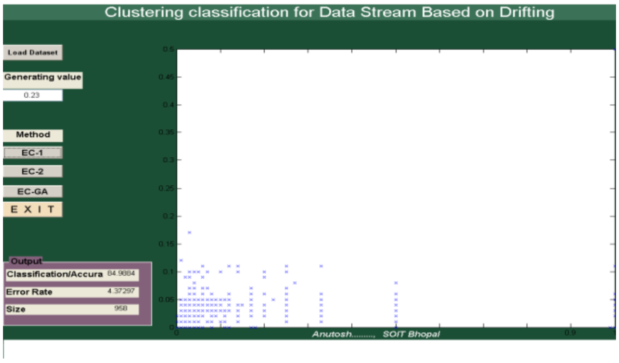
It means that combination of Genetic algorithm with classification and cluster technique enhances the accuracy of cluster and classification and also at the same time decreases the error rate. From figure (2), it is shown that randomly seed value is generated which must be in the data range. And after this, values nearer to the seed are calculated. This calculation is done for all the generated seed. Finally mean value is calculated. Since we are using concept of genetic algorithm in our work, so firstly fitness function is calculated for generated seeds. Each time mean value of generated population is compared with the fitness function. If the mean value is greater than the fitness function then only it is taken otherwise it is discarded. Output of this stage is the list of all the seeds which are satisfying the condition. After this stage we take the set of all the seeds as an input and generate clusters and classifies the data on the basis of these seeds.Here using of fitness function, is the only reason for better performance of our approach over the existing approaches.

5. EXPERIMENT RESULT

The above algorithms were implemented in Matlab 7.2 simulation tool.Tictoc dataset has been used to show the comparative result of two existing methods which are referred to as EC1 and EC2 and our newly proposed approach which is based on the concept of Genetic algorithm. It has been proved by our results that combination of Genetic algorithm with classifiers and clusters are improving the results.

Ensemble classifiers built on different data chunks using the same learning algorithm [9] (Denoted as EC1), and ensemble classifiers built with different learning algorithms on the up-to-date chunk [10] (denoted as EC2). In EC1, we use Decision Tree as the basic learning algorithm. In EC2, we use logistic regression classifier, decision tree, and Naive Baye's as the basic algorithms.

Here we have taken two data sets train data set and test data set. Train data set is the data set which has been trained by neural network approach. Whenever any input is given by us then it is compared with this train data set and each time distance is measured between the test data and train data. If the distance between these is in the under limit which has been specified by us in the algorithm then all the elements which falls under the limit will be grouped in the cluster. So each time when we enter a new generating value it is trained with the trained data and at this stage classifier works it means decision is made that whether this seed value is going to make a new cluster or it is present in the trained data cluster.Figure 3 shows the screen where dataset is loaded by clicking on the button Load Dataset and a random value between 0 and 1 entered as 0.23. The random value works as a seed or generating value for a cluster.


Clustering classification for Data Stream Based on Drifting
Load Dataset
Method
EC-1
EC-2
EC-GA
E X I T
Output
0.5
0.45
0.4
0.35
0.3
0.25
0.2
0.15
0.1
0.05
0
0
0.9
Anutosh........., SOIT Bhopal

Generating value 0.23

Classification/Accura 84.9884

Error Rate 4.37297

Size 958

Figure 3: Loading of dataset and insertion the seed value

Here we are taking Classification accuracy and Error rate as a comparative parameter with our method and the two existing methods.

Now we will determine Classification accuracy and Error rate with same seed value 0.23 for all three methods namely EC-1, EC-2, and EC-GA.

Figure 3 shows result corresponding to EC1 method. The value generated for classification accuracy is 84.9884 and Error rate which we get 4.37297, which is good but not at a satisfactory level.

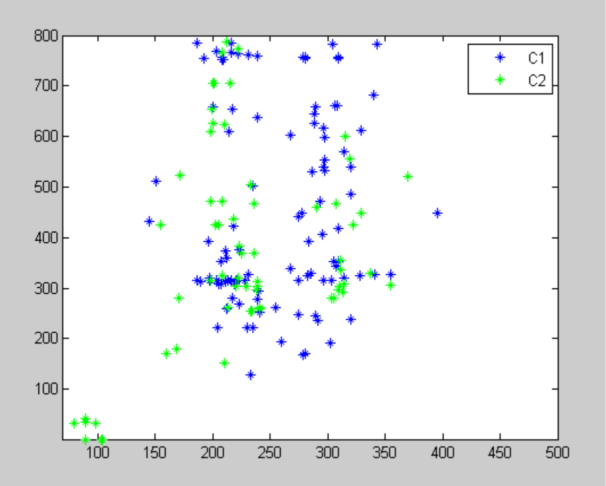


Figure 4: Result corresponding to EC1

Now we will check the same seed value 0.23 for the EC 2 method which is illustrated in figure 5. Classifier accuracy is 86.467 and Error rate is 5.94016. Although efficiency is too good comparatively but still Error rate is also increasing correspondingly. Figure 5 shows result corresponding to EC-2 method.

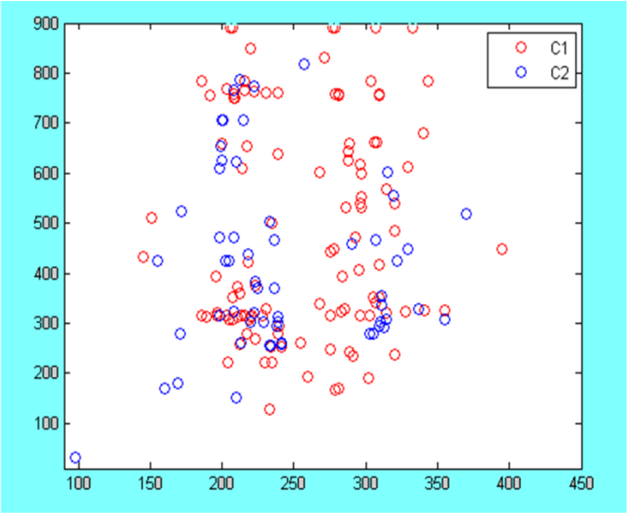


Figure 5: Result corresponding to EC2

Because of unsatisfactory previous results we applied a new approach, based on decision tree concept in combination with Genetic algorithm. For the seed value 0.23, Classification accuracy is 93.9765 and Error rate is 1.67577. Figure 6 shows result corresponding to EC-GA method.

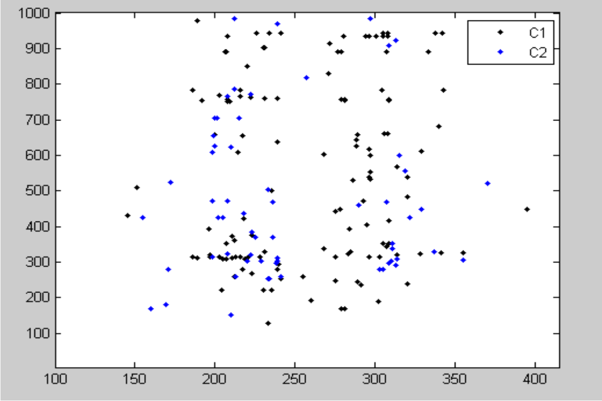


Figure 6: Result corresponding to EC-GA

The above result set shows that for the same value of seed the EC-GA have accuracy of 93.9765 which is greater than both the previous methods. Also the Error rate is reduced to 1.67577, which is a great improvement on the previous methods. Above results are shown using following table.

Table 1: Table of results of all three methods

SIZE – 958GV-0.23	EC1	EC2	EC-GA
Classification Accuracy	84.9884	86.467	93.9765
Error Rate	4.37297	5.94016	1.67577

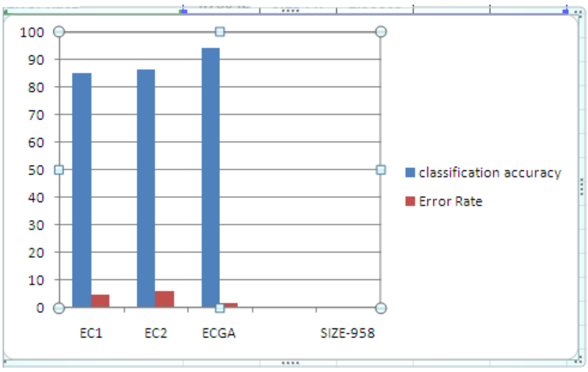


Figure 7: Comparative graph of EC1, EC2 and EC-GA

Also can be clearly understood by the graph generated by the classification accuracy and Error rate for all the three methods. This is illustrated in figure 7.

6. CONCLUSION

Clustering and classification ensemble learning is a commonly used tool for building prediction models from data streams, due to its intrinsic nature of handling large volumes stream data. In this research work we have presented a new ensemble method of clustering and classification using genetic algorithm to mine data stream. The contribution of this Paper is to introduce a new way of stream data mining, and to show how we can improve the accuracy and error rate, accomplished by new ensemble model with the use of genetic algorithm. The main reason for using genetic algorithms along with clustering and classification was its high ability to solve optimization.

Future work include some better tuning of Genetic algorithm with the classifier for achieving better results.Also, we can use another optimization process such as ANT, PSO and biological inspired function.ANT is known as ant colony optimizationalgorithm (ACO) is a probabilistic technique for solving computational problems. PSO known as particle swarmoptimization is a computational method that optimizes a problem byiteratively trying to improve a candidate solution with regard to a given measure of quality.

7. REFERENCES

[1] BPeng Zhang, XingquanZhu ,Jianlong Tan , Li Guo. 2007. Classifier and Cluster Ensembles for Mining Concept Drifting Data Streams. QCIS Centre, Faculty of Eng. & IT, Univ. of Technology.

[2] Reza Ghaemi, Nasir bin Sulaiman, Hamidah Ibrahim, Norwati Mustapha. 2010. A review: accuracy optimization in clustering ensembles using genetic algorithms Springer Science+Business Media B.V.

[3] Mohamed MedhatGaber, ArkadyZaslavsky and ShonaliKrishnaswamy. Mining Data Streams: A Review. Centre for Distributed Systems and Software Engineering, Monash University, 900 Dandenong Rd, Caulfield East, VIC3145, Australia.

[4] C. Shafer, R. Agrawal, and M. Mehta. 1996 Sprint: A scalable parallel classifier for data mining. In VLDB.

[5] J. Gehrke, R. Ramakrishnan, and V. Ganti. RainForest. 1998. A framework for fast decision tree construction of large datasets. In VLDB.

[6] J. Gehrke, V. Ganti, R. Ramakrishnan, and W. Loh. 1999. BOAT– optimistic decision tree construction. In SIGMOD.

[7] P. E. Utgoff. 1989. Incremental induction of decision trees Machine Learning, 4:161–186.

[8] P. Domingos and G. Hulten. 2000. Mining high-speed data streams. In SIGKDD, pages 71–80, Boston, MA, ACM Press.

[9] H.Wang, W. Fan, P. Yu, J. Han. 2003. Mining concept-drifting data streams using ensemble classifiers,In Proc. of KDD.

[12]

[10] J. Gao, W. Fan, J. Han. On appropriate assumptions to mine data streams: analysis and practice, 2003 In Proc. of ICDM.

[11] S. W. Mahfoud. 1993. Simple analytical models of genetic algorithms for multimodal function optimization. Proc. of 5th ICGA.

[12] S. W. Mahfoud. 1993. Cross over interaction among niches.Proc. of 1st IEEE Conference on Evolutionary Computation, world Congress on Computation Intelligence, 188-193, 1993.88-193.

[13] P. K. Nanda, D. P. Muni, P. Kanungo. 2002. Parallelized crowding scheme using a new interconnectionmodel. International Conference on Fuzzy Systems, Calcutta, LNAI(2275), Springer-Verlag:436-443.