Classification of Concept-Drifting Data Streams using Optimized Genetic Algorithm

E. Padmalatha
Asst.prof
CBIT

C.R.K. Reddy, PhD
Professor
CBIT

B. Padmaja Rani, PhD
Professor
JNTUH

ABSTRACT
Data Stream Mining is the process of extracting knowledge structures from continuous, rapid data records. In these applications, the main goal is to predict the class or value of new instances in the data stream given some knowledge about the class membership or values of previous instances in the data stream. Machine learning techniques can be used to learn this prediction task from labeled examples in an automated fashion.In many applications which are in non-stationary environments, the distribution underlying the instances or the rules underlying their labeling may change over time, i.e., the class or the target value to be predicted may change over time. This problem is referred to as Concept drift[8]. Evolutionary Computations like Genetic Algorithm is a strong rule based classification algorithm which is used for mining static small data sets and inefficient for large data streams. Evolutionary Algorithms are one of the population optimization techniques done by calculating fitness evaluation measures using gene reproduction, crossover, mutation and selection of the individual gene mechanisms. If the Genetic Algorithm can be made scalable and adaptable by reducing its I/O intensity, it will become an efficient and effective tool for mining large data sets like data streams.In this paper a scalable and adaptable online genetic algorithm is proposed to mine classification rules for the data streams with concept drifts. The results of the proposed method are comparable with the other standard methods which are used for mining the data streams.

Keywords
Data Stream, conceptdrift, Genetic Algorithm,optimization.

1. INTRODUCTION
Genetic Algorithm [1, 4and 6] is a rule based classifier whose performance will be almost similar to RBC. The Evolution of GA was started from the Darwinian's Theory "Survival of the fittest". It also has some major advantages over RBC. To make the classifier building process faster and easier, RBC stores a compressed form of the data stream in the memory as a tree. Since the stream evolves abruptly, frequent and fast modification of both the trees is also required. Hence, when the domain becomes too complex, building and maintaining the trees becomes a difficult task.

Compared to RBC, GA model is independent of the domain knowledge and does not require any complex data structures to store the data. So its memory requirement is low and does not require any complex computations as required for RBC. Due to its evolutionary based characteristic, it can handle the concept drifts in a natural way and the model can be made to evolve and adapt itself in accordance with the changes in the concepts of the data streams due to concept drifts. On the other hand GA scans the data set repeatedly to check the accuracy of the candidate rule set after each generation which is not possible with respect to the data streams, as the data streams cannot be accessed repeatedly. Hence here a scalable and adaptable GA is built for large data sets like data streams by reducing its I/O intensity.

2. RELATED WORK
Ensemble Classifiers (EC)

EC [3, 5] builds and uses a group of classifiers for predicting the class label of the new unknown data sample. In this type of algorithm, the data stream is divided into weighted chunks and classifiers are built for each chunk separately as shown in Fig1. The newly built last classifiers are used for predicting the new data samples. This collective decision making increases the prediction accuracy[7] when compared to the prediction accuracy of the models that employ only a single classifier for the prediction purpose.

Example for Ensemble Classifier is providing the same raw material to design a product to different designers according to their weightages based on their experience. Usually ensemble algorithms perform poorly while predicting the data samples of rare classes, particularly when the data distribution is highly skewed.

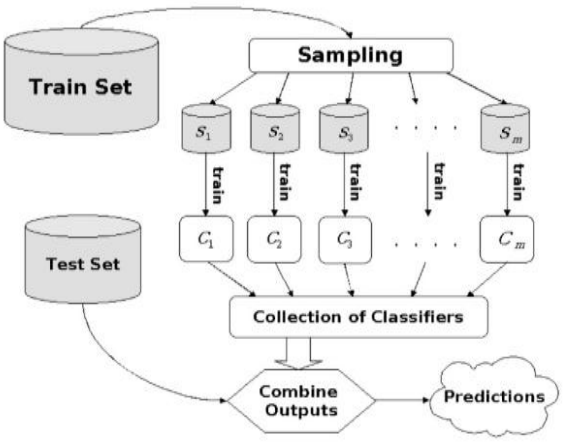

Figure 1. Ensemble Classifier

Rule Based Classifiers (RBC)
The third category of algorithms considers the classifier as a combination of tiny independent components and each component is built independently in an incremental way. The most recent algorithm of this type is low granularity Rule Based Classifier (RBC) proposed by Wang et al. 2007.Their way of building a classifier considerably reduces the model updating cost and their approach also maintains high accuracy level when compared to the first two approaches as their method upgrades only the components which are affected by the concept drift rather than making a global modification whenever a concept drift is detected, which is a faster and

easier process and makes the approach accurate and faster.

3. DESIGN OF OPTIMIZED GA

3.1 OGA Functionalities

Methodology of Optimized GA includes four functionalities shown in Fig2:

1. **Data stream distributor**
   Initially, training dataset and test datasets are taken. Here, test dataset is taken as input and training dataset is uploaded for streaming the data. Data stream distributor is responsible for streaming the uploaded training dataset continuously.

2. **Population creator**
   During the process of data streaming, Population creator creates initial population and individuals with duplicate instances in series of windows. Then these population generated windows are given to Genetic engine.

3. **Genetic engine**
   Then the Genetic engine performs OGA mechanisms on set of populations created by calculating the fitness values. Genetic Engine mechanisms include following steps like selection, crossover, mutation and elitism selection of individuals.

4. **Rule set Evaluator**
   Here, after calculation of fitness values of all the individuals, rules are generated for solution, i.e., genes values are calculated with their classified run time. Rule set evaluator generates the best rules with best fitness values after all the iterations.

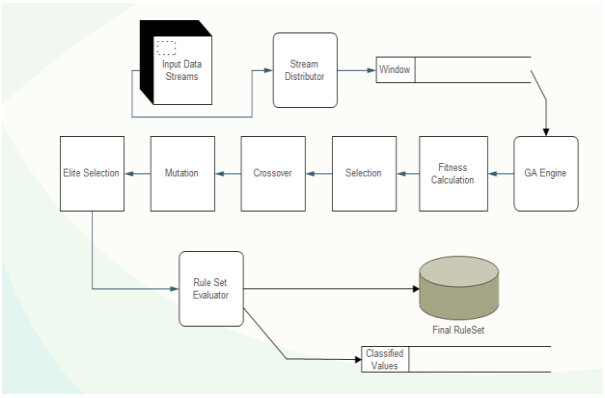

Figure 2. Design Flow of OGA Process

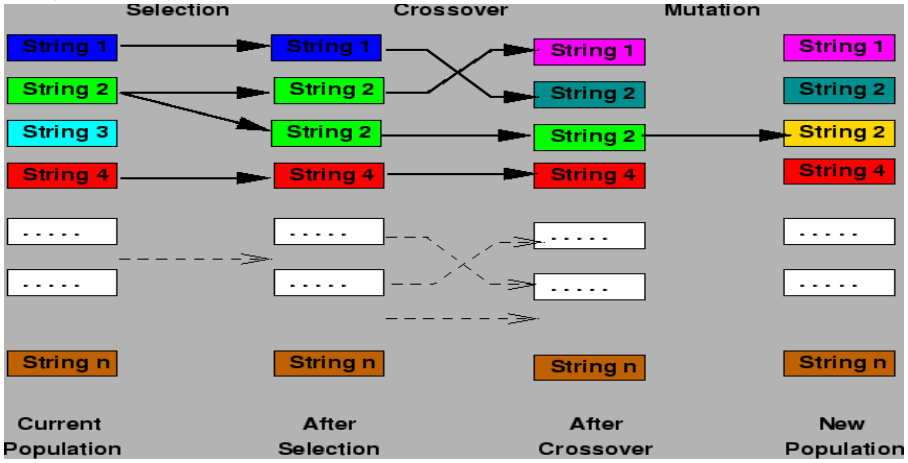

Figure 3 OGA Process

3.2 OGA Process with Datasets

Considering car data sets, which contain 1728 records and 6 attributes, all attributes are categorical. The target class attribute has four values namely 'unacc', 'acc', 'good', 'vgood'. To generate larger data sets of size 10000, 20000 and 30000 the records are duplicated and randomly arranged such that the data distribution is proportionately similar to the original data set.

Attribute	Values
Buying	vhigh, high, med, low
Maintenance	vhigh, high, med, low
Doors	2, 3, 4, 5more
Persons	2, 4, more
Lug boot	Small, med, big
Safety	Low, med, high

Now here in OGA Process,

Creation of Population is duplicating the records with size say suppose 1000 set of records from training data sets.

Individuals are the sets of records. Here in car data set, the example for individual is,

vhigh, vhigh, 3, 2, small, high, unacc

Chromosomes are the combination of target class and individual for generating the solution. Example for chromosome is,

target: acc and vhigh, med, 3, more, med, med

Genes are the solutions found after generating the solution in GA process with assigned target class label value. Example for genes isf`,

vhigh vhigh 3 2 small high unacc

Fitness value of an individual is the measure value of the fitness function for that individual. Here, fitness value is initiated with a minimum threshold value based on the best elitism selection.

Now, the OGA Process shown fig3 for car datasets is done in the following steps:

target class attributes unacc, acc, good and vgood is considered as 1000, 0100, 0010 and 0001.

Similarly for the other attributes

Attribute	Values
Buying	vhigh-1000, high-0100, med-0010 and low-0001
Maintenance	vhigh-1000, high-0100, med-0010 and low-0001.
Doors	2-1000, 3-0100, 4-0010 and 5more-0001
Persons	2-1000, 4-0100 and more-0010
Lug boot	Small-1000, med-0100 and big-0010
Safety	Low-1000, med-0100 and high-0010

Now for example, the individuals

vhigh vhigh 3 2 small high unacc Is considered as,

1000 1000 0100 1000 1000 0010 1000

Chromosomes are formed with the target class attribute for unacc-1000 and the individual for generating the solution. So total 7 attributes and 4 target class attributes forms 28 chromosomes.

Similarly, for all the rules genes solution set is generated.The same OGA process is applied for other datasets also.

4. EXPERIMENTATION AND RESULTS

Experimental Process

1. Initially, any dataset that has to be taken is divided into two datasets, training (80%) and test (20%) datasets.
2. Then, the test dataset is taken as input.txt file in OGA and the training dataset is to be uploaded.
3. Then the uploaded training data set is to be streamed. The OGA process starts on the streamed datasets.
4. The OGA Process generates the genes solution for the corresponding target class attribute.
5. After generating the solution, the OGA displays the correctly and incorrectly classified attributes.
6. Finally, the classified run time (in nano seconds) of the generated best genes are shown with the dataset record index size
7. Graph between Runtime V/s Solution Dataset size value is plotted as shown in the following figure

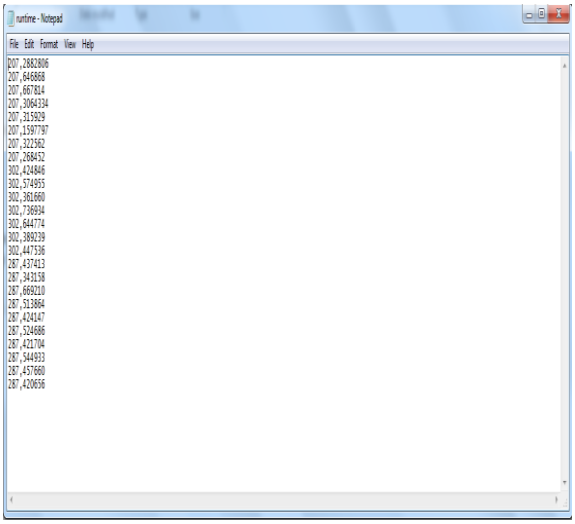

Fig4.Classification Run Time (in Nano Seconds) of OGA after generating the classified solution value for Yeast datasets

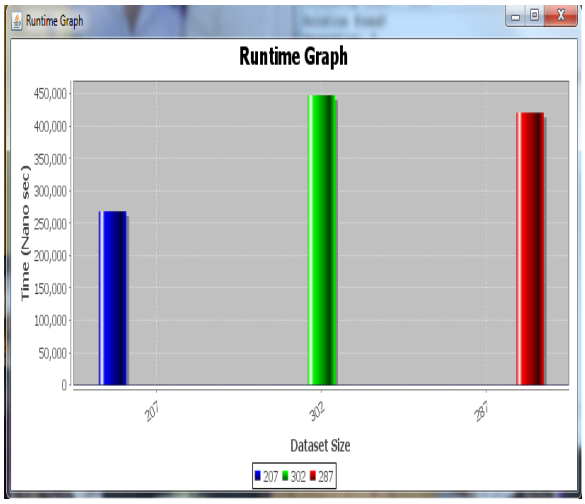

Fig 5. Run Time Graph of the Generated Classified Values of Yeast datasets

5. PERFORMANCE EVALUATION USING RIVAL ALGORITHMS

Considering

1. Error Rate which is equal to the ratio of incorrectly classified values and 100
2. Classification Run Time.

Table1. Classification Run Time (Seconds) tabulated using Different Classifications for 10 different datasets

Index	Dataset	EC (Random Forest)	RBC (PART)	CVFDT	Optimized GA
1	KDDCup	57.33	164.58	131.84	0.01811
2	Car	0.27	0.13	1	0.00084
3	Chess	5.03	39.52	2	0.001315
4	Nursery	6.021	0.15	2	0.00087

5	Hyperplane	7.43	0.14	2	0.015452
6	Sea	6.78	0.12	1	0.004276
7	Letter	26.7	0.1	1.5	0.01417
8	Image Segmentation	0.18	0.01	0.07	0.000724
9	Solar Flare	0.13	0.01	0.04	0.000892
10	Yeast Database	1.48	0.02	0.49	0.003023

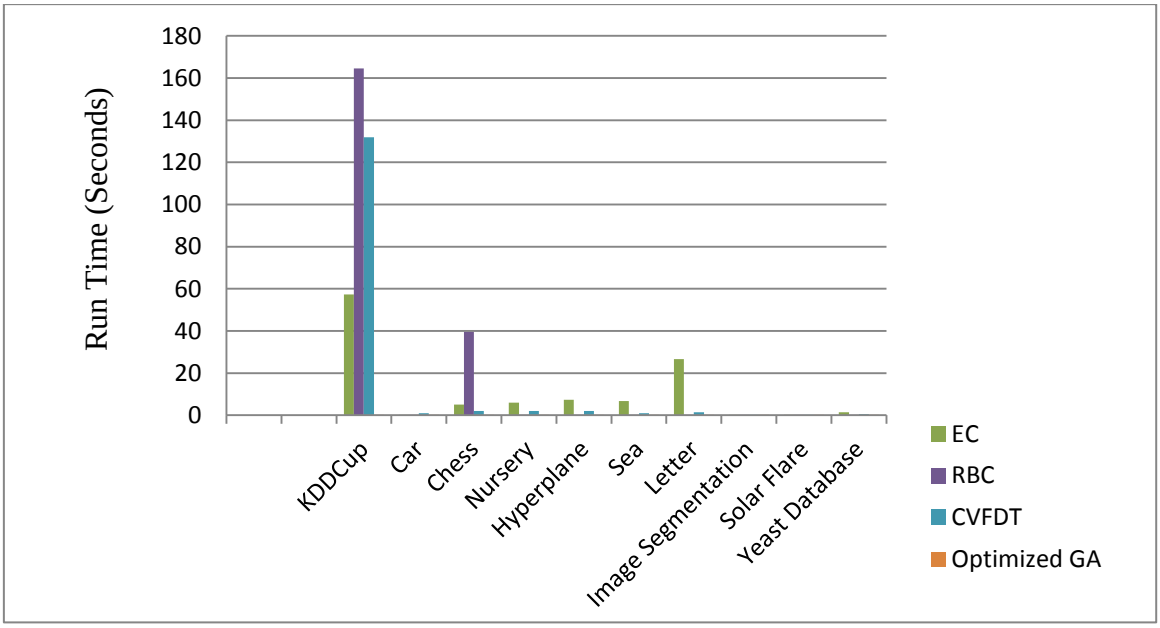

Fig 4.Comparison of Classification Run Time (Seconds) using Different Classifications for

Table 2. Error Rates tabulated using Different Classifications for 10 different datasets

Index	Dataset	EC (Random Forest)	RBC (PART)	CVFDT	Optimized GA
1	KDDCup	0.003	0.002375	0.15	0
2	Car	0.251	0.29978	0.29978	0
3	Chess	0.133412	0.122407	0.916844	0.001
4	Nursery	0.125	0.666667	0.498302	0.0015
5	Hyperplane	0.235	0.5378	0.4531	0
6	Sea	0.0895	0.4744	0.3741	0. 2
7	Letter	0.1389	0.98	0.603	0.49269
8	Image Segmentation	0.0435	0.95	0.0823	0.001
9	Solar Flare	0.168	0.1456	0.1183	0.001
10	Yeast Database	0	0.425202	0.38814	0

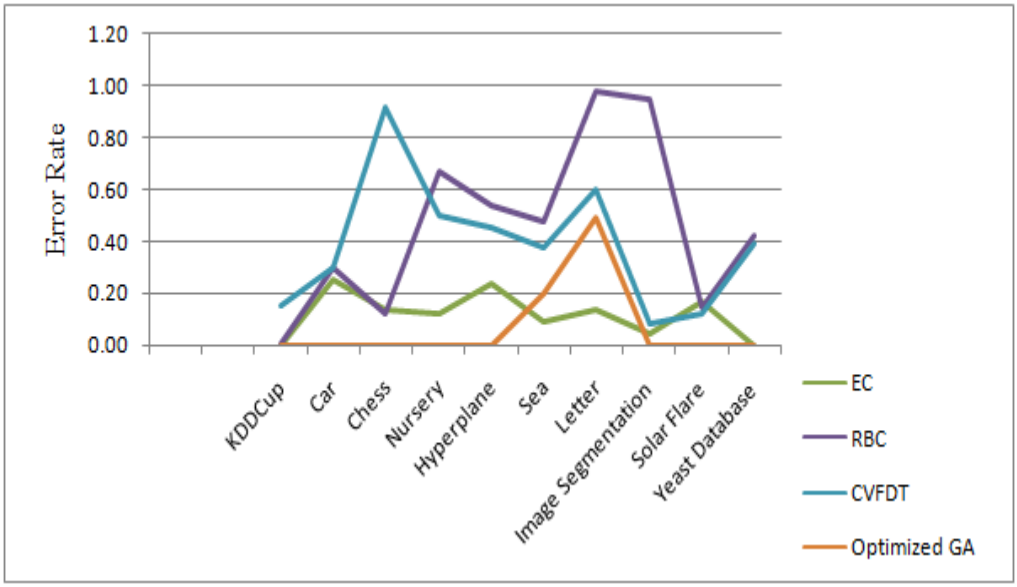

Fig6. Comparison of Error Rates using Different Classifications for 10 different datasets

6. RIVALS ALGORITHM

To compare the algorithms' performance, error rate and run time of data sets are calculated. A win/lose/tie (w/l/t) record is calculated for each pair of the method for which the experiment is performed.

It represents the number of data sets in which an algorithm, respectively wins, looses or ties when compared with the other algorithm regarding error rate. Same is calculated for all algorithms with respect to run time. From that we can prove which algorithm has best performance.

Table3. Performance Evaluation Using Rival Algorithm's w/l/t records with regard to their run time across 10 datasets

Method	EC	RBC	CVFDT	Optimized GA
EC	0/0/10	2/8/0	2/8/0	0/10/0
RBC	**8/2/0**	0/0/10	**8/2/0**	0/10/0
CVFDT	**8/2/0**	2/8/0	0/0/10	0/10/0
OGA	**10/0/0**	**10/0/0**	**10/0/0**	0/0/10

Table 4. Performance Evaluation Using Rival Algorithm's w/l/t records with regard to their error rates across 10 datasets

Method	EC	RBC	CVFDT	Optimized GA
EC	0/0/10	**7/3/0**	**9/1/0**	2/7/1
RBC	3/7/0	0/0/10	2/7/1	0/10/0
CVFDT	1/9/0	**7/2/1**	0/0/10	0/10/0
OGA	**7/2/1**	**10/0/0**	**10/0/0**	0/0/10

Hence Optimized GA has highest winning probability from both classification error rate and run time which proves the best efficiency.

7. CONCLUSION

Unlike the existing data sets classification algorithms like CVFDT, RBC, EC and Traditional GA, it is not possible to classify the data streams underlying with the mechanism

called concept-drift where the data streams are changed due to some underlying context changes. Also the data streams are not stored fully in any of the earlier classification techniques due to their concept drift.

Optimized GA is such a technique where the classification is done for concept-drifting data streams by using streaming window and its mechanisms like selection, crossover, mutation and elitism for the generation of the solution with best fitness value for best classification rate.

Further, the OGA can be optimized by minimizing the build time for construction of the model for even large data sets when streamed which enhances performance and time efficiency.

8. REFERENCES

[1] Periasamy Vivekanandan and Raju Nedunchezhian, "Mining data streams with concept drifts using genetic algorithm", Artificial Intelligence Review, Vol. 36, Issue 3, pp 163-178, Springer, October 2011.

[2] Araujo D.L.A, Lopes H.S, Freitas A.A, "Rule discovery with a parallel genetic algorithm", In Proceedings of IEEE systems, man and cybernetics conference, Brazil, 1999.

[3] Wang H, "Mining Concept-Drifting Data Streams", IBM T.J. Watson Research Center, August 19, 2004.

[4] Basheer M. Al-Maqaleh and Hamid Shahbazkia, "A Genetic Algorithm for Discovering Classification Rules in Data Mining", International Journal of Computer Applications (0975-8887), Vol. 41-No. 18, March 2012.

[5] Wang H, Fan W, Yu PS, Han J, "Mining concept-drifting data streams using ensemble classifiers", In Proceedings of the 9th ACM SIGKDD international conference on knowledge discovery and data mining, pp 226–235, 2003.

[6] Syed Shaheena and Shaik Habeeb, "Classification Rule Discovery Using Genetic Algorithm-Based Approach", NIMRA Institute, Department of CSE, IJCTT Journal, Vol. 4, Issue 8, pp 2710-2715, August 2013.

[7] E Padmalatha, C R K Reddy and Padmaja B Rani. Article: Ensemble Classification for Drifting Concept. *International Journal of Computer Applications* 80(11):33-36, October 2013.

[8] E.Padmalatha,C.R.K.Reddy, B.Padmaja Rani "Classification of Concept Drift Data Streams"In the proceedings of the Fifth International Conference on Information Science and Applications .ICISA 2014.IEEE PP291-295, 2014.