Volume 5, Issue 1, January 2015 ISSN: 2277 128X
International Journal of Advanced Research in Computer Science and Software Engineering
Research Paper
Available online at: www.ijarcsse.com

An Approach: Combining Clustering and Classification for Handling DataStream

Purva Gogte, Chandrayani Rokde, Ashwini Yerlekar, Shilpa Kamble, Sumedha Chokhandre
Assistant Professor, Department of Information Technology
Dr Babasaheb Ambedkar College of Engineering & Research, Nagpur, Maharashtra, India

Abstract— Traditional databases have been used in applications that require static data storage and comprises complex querying. But, Nowdays due to increase in use of Internet a huge continously changing data is generated everyday. This dynamically changing data are nothing but the Data stream. They are unbounded, continuous,usually come with high speed and have a data distribution that often changes with time. It has different issues such as memory management, Timeline Query and many more. Handling such issues are very necessary.There is need of handling data streams because data stream may be labeled or it may be unlabelled. Classification is supervised it can only handle labeled data. Thus, there is need of Hybrid Ensemble Classifier in which clustering and classifier are used combinely so that the labeled as well as unlabelled datastream both can be handled efficiently.This Paper describes an approach of new Ensemble Classifier that includes Clustering technique as well,So that the unlabelled data can be handled and inorder to increase the prediction accuracy.

Keywords- Clustering, Classification, DataStreams

I. INTRODUCTION

Traditional databases have been used in applications that require persistent data storage and complex querying. Usually, a database consists of a set of objects, with insertions, updates, and deletions occurring less frequently than queries.But, Currently the use of Internet is increasing day by day that leads to generation of huge dynamic data that are Datastream. Data Stream mining has recently emerged as a growing field of multidisciplinary research. It combines various research areas such as machine learning, artificial intelligence, statistics, automated scientific discovery data visualization, databases and high performance computing thus, Data stream classification has been a widely studied research problem in recent years. The dynamic and evolving nature of data stream requires efficient and effective techniques that are different from static data classification techniques. In recent years mining data streams in large real time environments has become a challenging job due to wide range of applications that generate boundless stream of data such as log records, mobile application sensors, emails, blogging, creditcard, fraud detection, medical imaging, intrusion detection, weather monitoring, stock trading, planetary remote sensing etc.

Clustering and Classification are the two methods that are used to find out the patterns from the huge amount of data stream. Clustering is the one of the most important task of the data mining. Clustering is the unsupervised method to find the relations between points of dataset into several groups. It helps in uncovering useful structures in data that were previously unknown. Classification is a data mining function that assigns items in a collection to target categories or classes. The goal of classification is to accurately predict the target class for each case in the data.Classification is supervised thus it is able to work only on labeled data set, On the other hand Clustering is Unsupervised approach means, in this classes are not predefined.The classifier are of mainly two types Single Classifier and Ensemble Classifier, Tradionally, Single Classifiers were used but, recently an ensemble classifier is used. An ensemble classifier is conventionally constructed from a set of base classifiers that separately learn the class boundaries over the patterns in a training set. The decision of an ensemble classifier on a test pattern is produced by fusing the individual decisions of the base classifiers. Ensemble classifiers are also known as multiple classifier systems, committee of classifiers, and mixture of experts.

II. LITERATURE SURVEY

Classification is supervised learning method as in this class labels are already defined. There are major two types classifiers.

i. Single Classifier: These are the tradional Classifier such as Naive Bayes, Nearest neighbor methods, and decision rules.
ii. Ensemble Classifier: An ensemble Classifier is formed by combining multiple classifiers.

2.1 Single Classifier:

VFDT [13] is Single Classifier proposed by Domingos and Hulten .It is a decision tree algorithm order. It helps to overcome the long training times issue. VFDT is used for the Real-Time Data Mining of Imperfect Data Streams in a

Distributed Wireless Sensor Network. The VFDT system constructs a decision tree by using constant memory and constant time per sample. VFDT algorithm is based on a decision-tree learning method combined with sub-sampling of the entire data stream. Advantage: VFDT requires less memory.Disadvantage: It has various drawbacks such as whenever the size of the training set is small; the performance of this approach can be unsatisfactory.

CVFDT [4] Concept Adapting very Fast Decission Tree was proposed by Hulten et al. CVFDT is used to handle concept drift issue by growing alternate trees and subtrees. This algorithm mines high-speed data streams under the approach of one-pass mining. Advantage: It is faster than VFDT.Disadvantage:It cannot handle the concept drift. CVFDT takes more space and the accuracy of this model is not greater than the best sliding window model.

Random forests (RF)[10] are nothing but the combination of tree predictor each tree depends on the values of a random vector sampled independently. Suppose we are having given a training set S. Build subset Si by sampling and replacement. Choose best split from random subset of F features .Make predictions according to majority vote of the set of trees.Disadvantage: Random Forests has drawback such as the feature selection process is not explicit. It has weaker performance.

Hoeffding option Tree [2] allows each training example to update a set of option nodes rather than just a single leaf. Like standard decision tree nodes it can split the decision paths into several subtrees. It makes a decision with an option tree by combining the predictions of all applicable leaves into a single result.Disadvantage: Hoeffding Option Tree is time consuming.

2.2 Ensemble Classifier:

A large amount of work has also focused on Ensemble Classifier such as Fast And Light Classifier, OzaBag, OzaBoost, OzaBagADWIN, OzaASHT. There are two major types of Ensemble Classifier that are Bagging and Boosting.Bagging [18] is sampling based approach has been proposed by Breiman. It generates multiple base classifiers by training them randomly.Finally the decission is taken according to majority voting. Boosting has been proposed by Schapire. It creates the data subsets for base classifier training by resampling the training data.

OzaBag [12] online bagging. It has been proposed by Oza and Russel.This is used for stream data classification.Online bagging is a good approximation to batch bagging.Disadvantage: OzaBag cannot handle gradual and sudden concept drift and it requires more memory space.

OzaBoost [12] is a online boosting algorithm. It generates a sequence of base models using weighted training sets and the correctly classified examples are given the remaining half of theweight.Disadvantage: Oza Boost cannot handle gradual and sudden concept drift.

OzaBagASHT [7] is new bagging method which use Adaptive Size Hoeffding Tree that sets the size for each tree.If the number of split nodes of the ASHT tree is higher than the maximum value, then it deletes some nodes to reduce its size. It is bagging using trees of different size.Disadvantage: OzaBagASHT cannot handle abrupt concept drift,

OzaBagADWIN [7] algorithm is to use a sliding window, not fixed a priori, whose size is recomputed in online according to the change rate observed from the data in window itself. Disadvantage: OzaBagADWIN cannot handle abrupt concept drift.

Fast and light classifier [6] is a new ensemble method for classification of stream data. It uses adaptive windowing technique for change detection and estimation and it uses the boosting technique with hoeffding tree for building ensembles.It also deals better with concept a drift which is crucial problem of evolving data streams. Disadvantage: Fast and light classifier is more accurate, in terms of time and memory in classifying both synthetic and real data sets.

2.3 Clustering Technique

There are many clustering techniques that were used such as DenStream, R-Den Stream, Density Grid Based and ClusStream**.**

Feng Cao.Et.al proposed Den Stream evolutionary algorithm. It is used for dealing with dynamic data stream. DENSTREAM is extension of DBSCAN. Disadvantage: Den Stream cannot distinguish clusters which have different levels of density and there is Loss of knowledge points.

In rDenStream [11] clustering, dropped micro-clusters are stored on outside memory temporarily, and new chance is given to attend clustering to improve the clustering accuracy. Disadvantage: rDenstream needs more memory space, because it needs external disk to memorize historical outliers

Density Grid Based [15] adopts a density decaying technique to capture the evolving data stream and extracts the boundary point of grid.It resolves the problem of evolving automatic clustering of real-time data streams.Density Grid Based has better scalability in processing large-scale and high dimensional stream data. Disadvantage:It cannot find arbitrary shaped clusters with noise.

CluStream [8] has been proposed by Aggarwal.In this method the clustering process is divided into two parts: online and offline.The online part clusters coming data divided by time window and store the results. The offline part generates the clustering results based on observation. Disadvantage: CluStream fails to handle changing data, thus leads to lost of knowledge point.

III. PROPOSED APPROACH

In this proposed approach we are combining Clustering as well as Classification together So, that Unlabelled datastream can also handled and prediction accuracy can be improved.The Forest cover Type dataset is taken as input

stream.The proposed Architecture is as given in Fig1.In this architecture The Forest Covertype DataSet is taken as a input Stream.Then Different Clustering Techniques are applied, after that N number of Classifiers are used. Forest Cover Type dataset is huge dataset that consist of near about 500000 entities. It contains seven Classes.

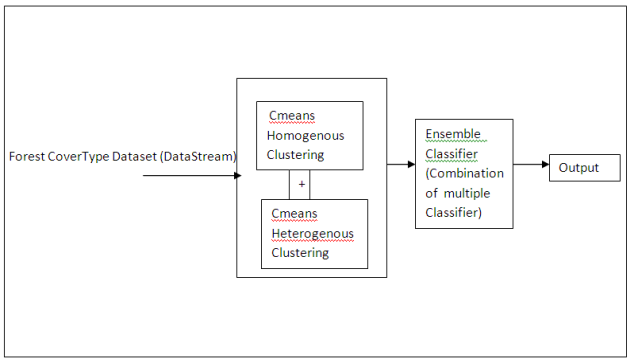

Fig: 1.Proposed architecture

3.1 Proposed Algorithm:

1. Input The DataStream { d1,d2.....den}
2. Divide The DataStream into n windows i.e. From{w1,w2....win} using Windowing Technique
3. Perform Clustering Using different Clustering Techniques.
4. Perform Classification using different Classifie**r**
5. Get the Output in terms of improved accuracy

The Following Techniques will be used So that the accuracy can be improved.

- **Cmeans heterogenous Clustering**: This is the Clustering Algorithm that basically forms the Clusters of elements that belongs to different Classes
- **Cmeans homogenous Clustering:**This is the Clustering Algorithm that basically forms the Clusters of elements that belongs to same Classes.
- **Naive Baysian Classifier: It** is a simple probabilistic classifier based on the Bayes' theorem with strong naive independence assumptions. It is used for calculating the number of the elements presents in each cluster. It basically demonstrates the amount of data present in each cluster.
- **Decission Tree:** Decision Tree Classifier is a simple and widely used classification technique. It applies a straitforward idea to solve the classification problem. Decision Tree Classifier poses a series of carefully crafted questions about the attributes of the test record. Each time it receives an answer, a follow-up question is asked until a conclusion about the class label of the record is reached.
- **Random Forest:** It is an ensemble learning method for classification and regression that operate by constructing a multitude of decision trees at training time and outputting the class that is the mode of the classes output by individual trees.

In This Project we are basically implementing an Ensemble approach so, that the accuracy can be improved and we are also using Clustering with the Classifier so, that the Learning Domain can be improved.

Algorithm: Ensemble Classifier Prediction

Step1: Input: test the pattern {e1, e2…en }

Step2: Output Cluster Confidence Vector {c1, c2…cn}

Step3: Compute Cluster Confidence Vector for each classifier

Step4: Compute Class confidence vector for each cluster

The project consist of following Modules

- Training Data
- Multiple Clusters
- Base Classifier
- Fusion Classifier
- Prediction

A) Training Data:

- Collect a number of bench mark datasets from UCI machine learning repository to verify the strength of COEC.
- The bench mark datasets is Forest Cover Type Dataset taken from UCI.
- Get the particular dataset and convert into matrix format.
- Get data matrix and class matrix.

B) Multiple Clusters

The training datasets are first clustered using multiple clusters algorithm with FUZZY clustering. We have used two types of clustering in COEC.

- Heterogeneous clustering for partitioning all the patterns in the training set independent of any knowledge of the class of the patterns.
- Homogeneous clustering for partitioning the patterns belonging to a single class only. Patterns belonging to each class are partitioned separately.

C) Base Classifier

A set of base classifiers are trained with and produced by the clustering algorithm. The input to each base classifier is set to [dij]. The target for each base classifier is set to [tik] such that

tik = 1, if clusteri = k

tik = 0, otherwise where $1 \leq k \leq N\text{clusters}$.

D) Prediction

Get the training data and convert into test pattern $e = \langle e_1, \ldots, e_{N\ features} \rangle$. Each base classifier b produces $N_{clusters}$ different confidence values $\langle w^b_1, \ldots w^b_{N\ clusters} \rangle$ that indicate the possibility of the pattern belonging to the different clusters.

IV. RESULTS & DISCUSSION

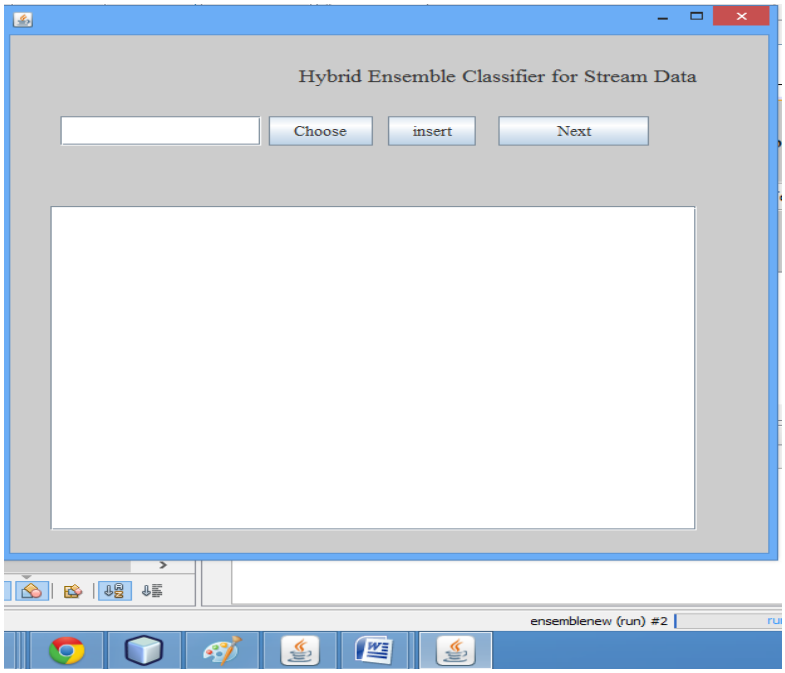

Fig.4. 1 Hybrid Ensemble Classifier for Stream Data

This Window consist of three buttons that are Choose, Insert and Next ButtonChoose Button is used for selecting the particular dataset. Insert is used for Inserting dataset Next Button is used for switching into the next window.

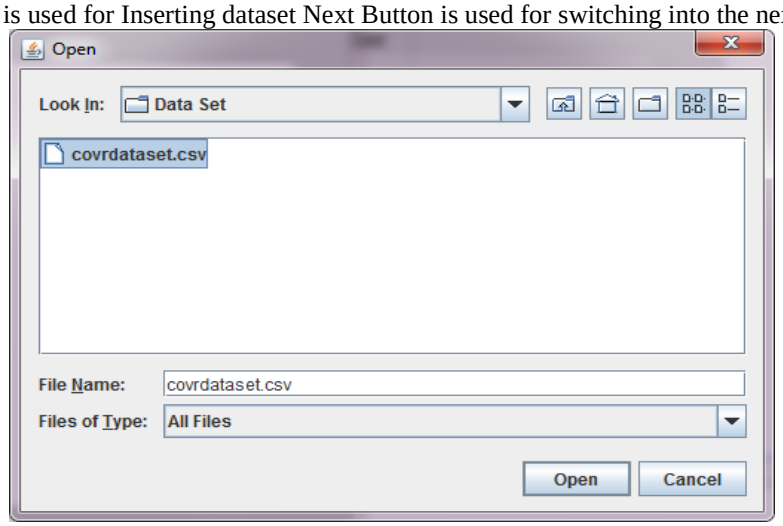

Fig.4.2 Dataset

This windows shows selection of particular dataset i.e Forest Cover Type dataset. The Forest Cover type dataset is the classification problem. There are Seven Classes in this data set. The actual Forest Cover type for a given observation is 30 x 30meter cell was determined from US Forest Service Data is in raw form and contains binary columns of data for qualitative independent variables. It consist of 500000 of instances 12 attributes.

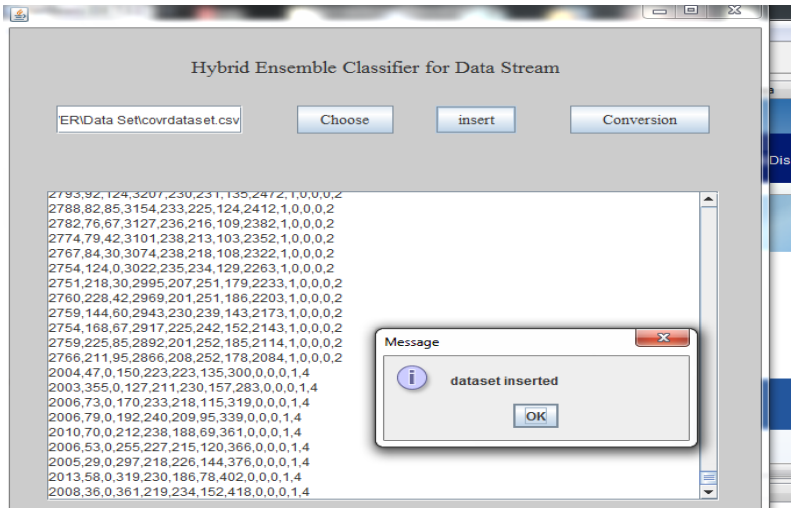

Fig.4.3 Forest Cover Type dataset

This Window shows the insertion of particular dataset i.e Forest Cover type and it's Connectivity with the database. There are Seven Classes in this data set. The actual Forest Cover type for a given observation is 30 x 30meter cell was determined from US Forest Service Data is in raw form and contains binary columns of data for qualitative independent variables.

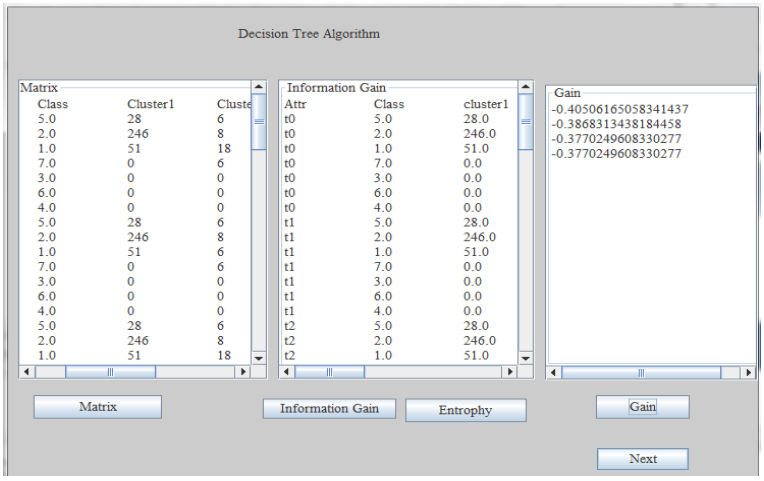

Fig.4.4 Decision tree

A decision tree is a tree in which each branch node represents a choice between a number of alternatives, and each leaf node represents a decision. Decision tree are commonly used for gaining information for the purpose of decision making. This window consists of three buttons. Matrix Button, Information Gain, Entropy and Gain. The Matrix Button Shows the data in the form of confusion Matrix .i.e. No of elements present in each class and the cluster are shown in matrix form. This confusion matrix is used in ID3 algorithm as a input.

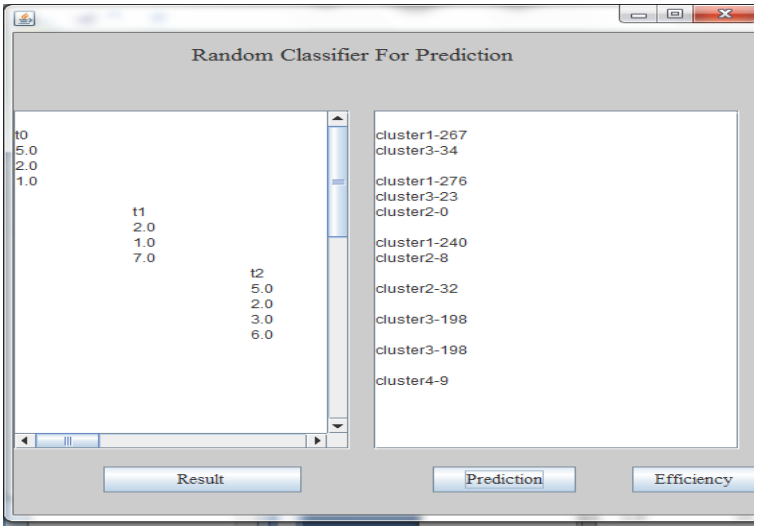

Fig.4.5 Random Base Classifier

In this window there are three buttons . Result ,Prediction and Efficiency button.The Result button shows the output of Random Classifier i. e it outputs the class that is the mode of the classes output by individual trees. Prediction Button indicates the number of patterns belonging to different clusters.

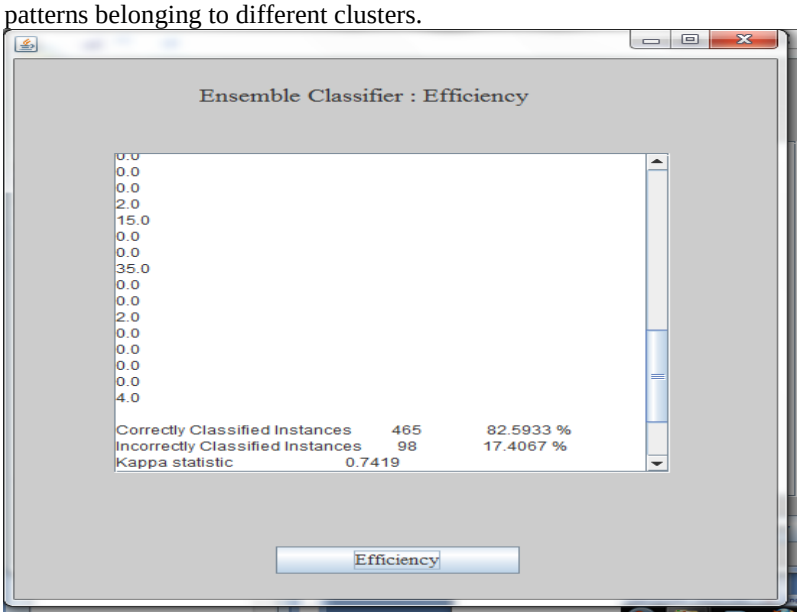

Fig.4.6 Efficiency of Ensemble Classifier

In this window there is button called efficiency .This button shows the efficiency of the hybrid approach. Efficiency is the number of instances correctly classified. Accuracy is calculated by finding the number of instances classified from the total no of instances.

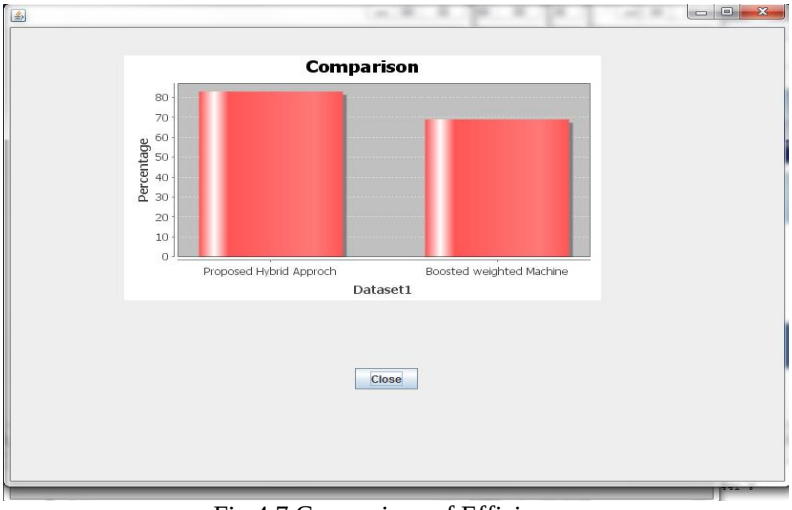

Fig.4.7 Comparison of Efficiency

In this window the comparison between the two techniques are shown .We have compared the accuracy of the Hybrid approach with the Boosted Weighted Machine. Accuracy is calculated by finding the number of instances classified from the total no of instances. We have compared the result of our approach with the Boosted Weighted Machine. The efficiency of our approach is 82% and that of Boosted weighted machine is 69%

V. CONCLUSIONS

Thus, there are different techniques that can handle the data stream but none of the clustering technique can handle the concept drift and as the datastream is huge and changing continuously it may be labeled or unabelled thus, there is need of an approach that can also handeled unlabelled stream so that datastream can handled more efficiently.

REFERENCES

[1] Brijesh Verma and Ashfaqur Rahman,"Cluster-Oriented Ensemble Classifier: Impact of Multicluster Characterization on Ensemble Classifier Learning, IEEE Trans.on Knowledge And Data Engineering, vol.24,pp.1156-1167, 2012.

[2] P.Chaudhuri A.K.Ghosh and H.Oja,"Classification Based on Hybridization of Parametric and Non-Parametric Classifiers,"IEEE Trans. Pattern Analysis and Machine Intelligence, vol.31,pp. 1153-1164, July 2009.

[3] Reza,O.Pujol,D.Masip,"Geometry Based Ensemble: Toward Structural Characterization of the Classification Boundary,"IEEE Trans. Pattern Analysis and Machine Intelligence, vol. 31, no. 6,pp. 1140-1146, June 2009 .

[4] Geoff Hulten, Laurie Spencer, Pedro Domingos,"Mining time changing data streams", ACM, USA, 97-106, 2001.

[5] Bieft A, Holmes G, Pfahringr B, Kirkby R,Gavalda R 2009,"New ensemble methods for evolving data streams." In KDD, pp 139–148, 2009.

[6] Kapil Wankhade,Vahida Attar,Pradeep Sinha,"A fast and light classifier for data streams", Springer, pp 200-207,2010

[7] Bifet A, Gavald'an R, Learning from time changing data with adaptive windowing. In: SIAM Int Conf Data Mining, pp 443–448,2007

[8] C. C.Aggarwal,J. Han, J. Wang, and P. S. Yu."A framework for clustering evolving data streams".In Proc. Of VLDB,pp. 81-92,2003.

[9] Feng Cao et al, Martin Ester, Weining Qian, and Aoying Zhou.Density-based clustering over an evolving data stream with noise. In SDM, 2006.

[10] L. Breiman, "Random Forests," Machine Learning, vol. 45, no. 1, pp. 5-32,Oct. 2001.

[11] HUANG Hai, LIU Li-xiong, rDenStream,A Clustering Algorithm over an Evolving Data Stream", The National High Technology Research and Development Program("863" Program), 2009.

[12] Oza N, Russell S "Online bagging and boosting." In: Artificial intelligence and statistics, Morgan Kaufmann, pp 105–112, 2001.

[13] P. Domingos and G. Hulten"Mining High-Speed Data Streams",In Proceedings of the Association for Computing Machinery Sixth International Conference on Knowledge Discovery and Data Mining,2000.

[14] Oza N, Russell S."Experimental comparisons of online and batch versions of bagging and boosting", In:ACM SIGKDD,pp 359–364,2001

[15] Amineh Amini, Teh Ying Wah, Mahmoud Reza Saybani, Saeed Reza, Aghabozorgi Sahaf Yazdi,"A Study of Density-Grid based Clustering Algorithms on Data Streams", Eighth International Conference on Fuzzy Systems and Knowledge Discovery (FSKD),2011.

[16] T. Windeatt, "Accuracy/Diversity and Ensemble MLP ClassifierDesign,"IEEE Trans. Neural Networks, vol. 17, no. 5, pp. 1194-1211,Sept. 2006.

[17] G.M. Munoz, D.H. Lobato, and A. Suarez,"An Analysis of Ensemble Pruning Techniques Based on Ordered Aggregation,"IEEE Trans. Pattern Analysis and Machine Intelligence, vol. 31, no. 2,pp. 245-259, Feb. 2009.

[18] L. Breiman, "Bagging Predictors," Machine Learning, vol. 24, no. 2,pp. 123-140, 2000.

[19] G. Fumera, F. Roli, and A. Serrau, "A Theoretical Analysis of Bagging as a Linear Combination of Classifiers,"IEEE Trans.Pattern Analysis and Machine Intelligence, vol. 30, no. 7, pp. 1293- 1299, July 2008.

[20] R.E.Schapire,"The Strength of Weak Learnability,"Machine Learning, vol. 5, no.2, pp. 197-227, 2000.

[21] R.E.Banfield,L.o.Hall,K.W.Bowyer,W.P.Kegelmeyer, "A New Ensemble Diversity Measure Applied to Thinning Ensembles,"Proc.Fourth Int'l Workshop Multiple Classifier Systems (MCS '03), pp. 306-316, 2003.

