ISSN:1991-8178

Australian Journal of Basic and Applied Sciences

Journal home page: www.ajbasweb.com

Genetic Based Ensemble Data Stream Classifier

[1]S. Jayanthi and [2]B.Karthikeyan

[1]*Research Scholar, Department of Computer Science and Engineering, Karpagam University, Coimbatore, Tamilnadu, India.*
[2]*Prof/Principal, Dhanalakshmi Srinivasan Institute of Research and Technology, Siruvachur, Perambalur, Tamilnadu, India.*

ARTICLE INFO

Article history:
Received 23 June 2015
Accepted 25 August 2015
Available online 2 September 2015

Keywords:
Concept drift, data stream classification, Neuro-Fuzzy, Genetic K-Means, concept evolution, ensemble.

ABSTRACT

In recent years, data stream classification has become an active research area as the need for analyzing widespread availability and manipulation of data streams has become indispensable. Concept drift and concept evolution are the most inevitable phenomenon that tone down the performance of the data stream classification task, and for which an acute solution has not yet been discovered. Concept drift occurs when the nature of data distribution centre and the definition or interest behind the target class changes over time. Concept evolution occurs at the emergence of novel classes. In light of addressing these issues and the limitations faced by the existing approaches, this research work expounds a novel Genetic based Ensemble Data Stream Classifier (GEDSC) which synergizes both the supervised and unsupervised classification algorithms. The proposed system analyzes if and how Neuro-Fuzzy supervised classification algorithm, and Genetic K-Means unsupervised algorithm can be deployed together to confront both the concept drift and the concept evolution plights together effectively. The proposed model is experimentally tested with a dedicated server by generating a real time dataset for video streaming application and its efficiency has been compared with the state of the art data stream classification algorithms in terms of various parameters.

© 2015 AENSI Publisher All rights reserved.

To Cite This Article: S. Jayanthi and B.Karthikeyan, Genetic Based Ensemble Data Stream Classifier. ***Aust. J. Basic & Appl. Sci.,*** *9(27): 409-415, 2015*

INTRODUCTION

Advancements in computing and communication technologies lead to the generation of insurmountable stream of data in almost all digitized organizations. Data stream mining is a contemporary research domain concerned on the analysis of these massive and dynamic data streams. Among various data stream mining tasks, data stream classification is the most prominent task that classifies the class labels by formulating a classification model (Albert Hung-Ren Ko *et al*, 2013, Jayanthi *et al*, 2014, 2015). Most of the existing classification methods function well only on static data sets by making sufficient number of scanning. Typically, data generated for an online classification process is subject to change over time due to the changes in distribution center and the interest behind the target concept which is popularly known as concept drift that makes the current classification model detrimental and stale. Hence, the classification model should be updated to learn the new concept so as to remain alive (Liang *et al*, 2012, Brzezinski *et al*, 2014). Since the data stream classification process is constrained with limited resources and scanning, achieving sustained accuracy amid of concept drift is a notable challenge (Bifet *et al*, 2009, Aggarwal *et al*, 2006, Tsai *et al*, 2008). Concept evolution is also an equivalent plight as concept drift that occurs in data stream classification (Masud *et al*, 2011, 2013).

In light of addressing these issues and the limitations faced by the existing approaches, this research work has expounded a novel genetic based ensemble data stream classifier which synergizes both the supervised and unsupervised classification algorithms to pep up the current state of data stream classification task. In this research work, Neuro-Fuzzy classifier and Genetic K-Means classifier are chosen as the instance of supervised classifier and unsupervised classifier respectively.

This classifier hybridation has not been found in any existing literature to confront the concept drift and concept evolution in real time video stream mining application.

However large number of research work has been attempted to find a solution for these issues, most of them was carried out with synthetic datasets where the attributes of a class, distribution center is not concealed. Hence it is obvious that the algorithm which works well on synthetic data sets may not be

Corresponding Author: S. Jayanthi, Research Scholar, Department of Computer Science and Engineering, Karpagam University, Coimbatore, 641 021, Tamilnadu, India.
Tel: +91 9965361764 E-mail: nigilakash@gmail.com

good in handling real time data streams where the attributes of a class generated by different distribution centers over time is obfusticated. Hearkening to this, the proposed work is employed in a real time environment to assert its efficiency in terms of accuracy over other algorithms.

This paper is organized into five segments. Following the segment of introduction, state of the art of data stream classification is addressed in segment two. Genetic Based Ensemble Data Stream Classifier Phenomena is discussed in section three. The system design and the experimental setup of the proposed research work are illustrated in segment four. In fifth segment, performance of the proposed system is accentuated after a thorough performance evaluation. Last segment is concluded with the directions for the future enhancements of the proposed system.

Related Work:

This section discusses about the state of the art of data stream classification methods. Researches carried out on this domain are categorized into three major categories, namely, online incremental learning approaches, block based learning approaches and ensemble learning approaches (Zang *et al*, 2014). Incremental learning approaches classify the data streams by incrementally adapting the classification model; it is observed that its efficacy is high only in confronting gradual concept drift. Batch learning classifies the data streams by grouping datasets into chunks of equal size with sliding window technique; it is observed that its efficacy is high only in addressing recurrent concept drift. The third category, ensemble learning approach performs classification by grouping the results of its component classifiers. This method is widely recognized as a most accurate classification model and also has been widely adopted due to its modularity (Bifet *et al*, 2009, Farid *et al*, 2013).

It is notable that the proposed work is centered on the ensemble learning as it aims at improving the accuracy of the classification task. Concept evolution is also a notable plight which occurs when new class or category emerges in the data stream.

Decision tree is the most popular static classification algorithm which classifies the datasets by calculating the best splitting attribute for each branch of the tree. Its simplicity leads to the development of many decision tree based data stream classification algorithms, such as Very Fast Decision Tree (VFDT), Concept adapting Very Fast Decision Tree (CVFDT), Flexible Decision Tree (FlexDT), etc. However decision tree is followed as a pioneer, it has a main drawback that its accuracy is detroited even for a small concept drift (Domingos *et al*, 2000, Liang *et al*, 2012, Yang *et al*, 2013).

Brzezinski *et al*, 2014, proposed an Online Accuracy Updated Ensemble (OAUE) algorithm aimed at combating against concept drift by combining both the incremental and block based learning algorithms. It implements windowing strategy to train, evaluate its component classifiers. Its drawback is that its accuracy level fluctuates between different datasets. Multi Class Miner (MCM) which has been proposed by Masud *et al*, 2014, also aimed at combating concept drift. It implements Gini Coefficient, K-means clustering, Adaptive Threshold techniques to address various issues in data stream classification. From the literature study, it is observed that MCM actively identifies novel classes, but it is neither dynamic and nor robust in confronting concept drift.

ECSMiner is an ensemble classification algorithm which aims at discovering novel classes, in the presence of concept drift. This algorithm employs decision tree and K-Nearest Neighbour algorithm as component classifiers along with silhouette coefficient to discover novel classes (Masud *et al*, 2013). Its accuracy level drops down when the concept drift level is high. ECM-BDF is also an ensemble based data stream classification model that works based on decision-feedback. It combats against novel classes and concept drift by applying maximum consensus of ensemble and unsupervised classification models (Liu *et al*, 2014).

Since the research work was carried out and proven on synthetic datasets and it might not be good at real time dataset. CDR-Tree is a rule mining tree aimed at discovering rules of concept drift. CDR-Tree is good at confronting concept drift only when the data blocks size is almost two. When the block size increases, CDR-Tree becomes more complex (Tsai *et al*, 2009).

Herein the few of the observations made across the literature survey is listed,

- Very few researches have been conducted on real time environment.
- Accuracy level of the data stream handling algorithms is fluctuated between the occurrences of different kinds of concept drifts.
- Many algorithms are good at handling particular type of concept drift.
- Finally, the algorithms which are efficient in handling various kinds of concept drifts are inattentive on novel class detection.

These observations instill towards the development of the proposed system

Genetic Based Ensemble Data Stream Classifier Phenomena:

This section highlights the uniqueness and importance of creating genetic based ensemble data stream classifier. It is known that Fuzzy logic is good at interpreting the classification rules, but lagging at generalizing the rules, and in converse, neural network is good at generalizing the classification rules and but not so in interpreting the rules. In this research, Back propagation neural network is chosen due to its accuracy and versatility. The ensemble of

modified Neuro-Fuzzy classifier has been formed to achieve both of its benefits that is, the good interpretation of fuzzy logic and good generalization of neural network while leaving its demerits. The modified Neuro-Fuzzy classifier classifies the data stream in case of no concept drift.

Genetic algorithm is good at producing optimized results by implementing crossover, mutation and fitness function. K-means algorithm is a popular distance based unsupervised classification algorithm, it classifies the data by finding interclass and intra class similarity between the instances of the classes and is calculated by forming similarity matrix. Its drawback is that is fixing center tendency which varies over time.

Hence the ensemble of modified Genetic based K-Means algorithm is formed to address concept evolution effectively while utilizing its merits and leaving its demerits. The proposed system, GEDSC is formed by synergizing the ensemble of modified Neuro-Fuzzy classifier and Genetic-K-Means algorithm to confront the issues in Data stream classification.

System Design:

This section illustrates the system design and the process involved in data stream classification task with Figure.1. Increasing number of data being escalated from clients is sliced into data chunks and is processed using sliding window technique which is a widely adopted method in many of the existing research work. Role of the proposed Genetic Based Ensemble Data Stream Classifier is twofold.

At first, the data chunks in the sliding window are inspected for classification by the ensemble of modified Neuro-Fuzzy classifier. That is, under normal circumstances, the ensemble of modified Neuro-Fuzzy classifier is applied for data stream classification. If any concept drift is found and also all instances are labeled, then the classification model is adapted to perform better classification in future.

Secondly, if any instance is found unlabeled, then it might lead to concept evolution. In such case, the ensemble of modified Genetic K-Means unsupervised classifier model is applied for detecting the novel class occurrence. During clustering of unlabeled instances, if interclass similarity between the instances of successive data is high, then novel class occurrence is affirmed, and the detail is immediately sent to Neuro-Fuzzy classification algorithm to train the classification model. If interclass similarity is not up to the threshold level, then the data are kept in a buffer to find interclass similarity with future data. If no relevance is found with future data, it is confirmed as outlier and is flushed out of the buffer.

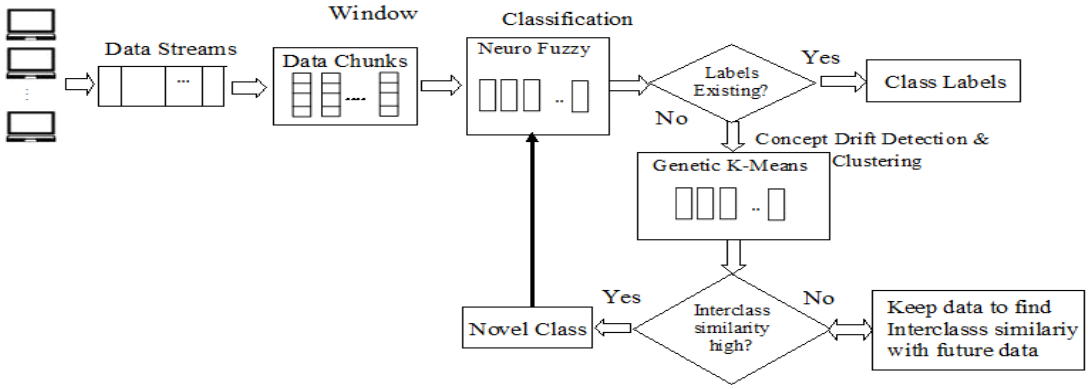

Fig. 1: Genetic Based Ensemble Data stream Classification Method.

Experimental Setup:

Hardware and Software:

To implement the proposed system, Enterprise STD dedicated server having the configuration IBM/Super/HP DL 160 G8 Series, 1 X Intel Hexa Core Xeon Processor E5-2620, 15 M Cache, 2.0 GHz ,7.2 GT/s Intel QPI), 2x Gigabit Ethernet Card is used for 6 months period to run our video stream mining applications. This model is tested by generating video requests from 25 personal computers having the configuration, windows 8.1(64 bit) operating system, Intel core I5 processor and Visual Studio 2013. However the request for videos can be generated from any type of the devices, such as, mobile, laptop, etc. A video emulator, ISVPFin has been developed to play videos on client side as shown in Figure.2.

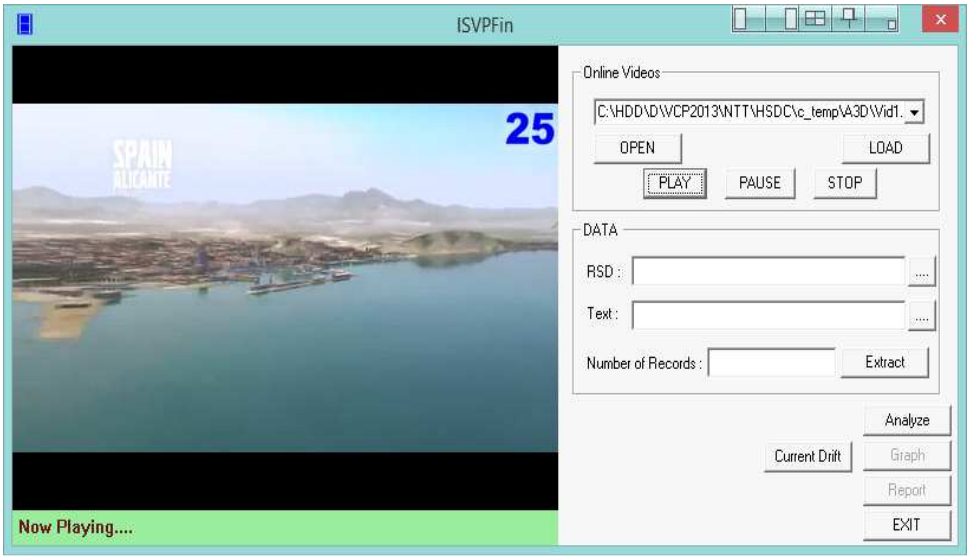

Fig. 2: ISVPFin Video Player.

Performance Evaluation:

User interest on videos are collected by a Real time dedicated server to prebuffer the anticipated videos in advance to reduce the video playout time, viewing latency and play out distortion. Real time server considers 16 attributes of the user demands and the corresponding 12 classes namely, Do It Yourself, Drama, Sport, Movie, Funny, Technology, 2D-Animation, 3D-Animation, Entertainment, Tutorials, Short Film and Vector. When there is a significant change found among 16 attributes such as, protocol, user name, IP address, source data header, destination data header, login status, number of failed logins, service count, type, total data size, duration, resolution, bit rate, repeat status, and audio quality, it is considered as concept drift.

Notable challenges in video stream mining process is twofold,

- Firstly, user interest on videos is dynamic which leads to concept drift and
- Secondly, the release of new category of videos is also very common which might lead to concept evolution.

If the user is not active for more than 30 days, then the user is considered as an inactive user and will be removed from the classifier to speed up the learning process so as to service the active users in a better way. The snapshot of dataset is shown below in Table 1. Most recent data stream classification methods, Online Accuracy Updated Ensemble (OAUE), Mine Class Method (MCM), ECSMiner, Ensemble Classification Model Based on Decision-Feedback (ECM-BDF), CDR-Tree are taken for comparing the results.

Performance evaluation is done in two parts:

1. By varying the size of the data chunks in sliding window, an exhaustive study on different contemporary algorithms are carried out. The performance of GEDSC algorithm on varying size of data chunk is slated in the table 2.
2. Performance of all studied methods is compared on various parameters and the average performance measure is slated in the table 3.

The performance of the studied data stream classification algorithms are illustrated using graphical plots as shown in Fig.3, Fig.4, Fig.5, Fig.6, Fig.7 and Fig.8 where horizontal axis represents number of data chunks and vertical axis represents percentage of measure for accuracy, precision, recall, F1-Measure, byte value for memory consumption and microseconds for processing. Through the research work carried out, it is observed that the performance of the proposed method is the most accurate and is measured using accuracy, precision, recall and F1-measure.

Accuracy is computed by using the formula:

Accuracy= (True Positives + True Negatives) / (True Positives + True Negatives + False Positives + False Negatives)

However accuracy measure ensures the accuracy of the classifier, accuracy measure is alone not sufficient to measure the correctness of a classifier. So precision, recall and F1-measure have also been calculated to strongly ensure the accuracy or correctness of the classifier.

Table 1: Snapshot of datastream

Protocol	IP Address	Source Data Header	Destination Data Header	Service Count	Type	Class
SFTP	144.94.37.237	4709	4410	119	MP4	Do It Yourself
IP	173.38.120.203	4707	3281	103	MPG	Drama
HTTP	158.4.234.31	2217	4627	60	MPG	Sport

FTP	230.124.207.249	1869	4162	13	WMV	Movie
FTP	194.78.58.110	2036	1487	57	MPG	Funny
HTTP	174.46.139.131	977	3892	1	WMV	Technology
FTP	143.1.163.225	977	3892	5	MPEG	2D-Animation

Report has been generated by GEDSC App on 26-07-2015 at 09:50:41

Table 2: Instance based Performance Report on GEDSC algorithm.

Procedure	Accuracy (%)	Precision (%)	Recall (%)	F1-Measures (%)	Memory (Bytes)	Processing Time (mS)	Avg.Response Time(mS)
10000	83.900000	83.950000	83.100000	83.522837	6285600	1498	977
20000	87.270000	87.560000	87.290000	87.424792	6446054	1656	981
30000	87.350000	87.620000	87.420000	87.519886	6450951	1807	1013
40000	87.550000	87.750000	87.650000	87.699971	6444586	1967	1003
50000	87.610000	87.820000	87.630000	87.724897	6445273	2120	1005
60000	87.660000	87.800000	87.690000	87.744966	6453240	2271	1010
70000	87.770000	88.010000	87.730000	87.869777	6453639	2419	1038
80000	87.850000	88.110000	87.950000	88.029927	6454736	2575	1081
90000	87.960000	88.140000	87.970000	88.054918	6453122	2724	1078
100000	87.930000	88.180000	87.950000	88.064850	6455835	2862	1073

Table 3: Average performance report of GEDSC and other comparative algorithms

Procedure	Accuracy (%)	Precision (%)	Recall (%)	F1-Measures (%)	Memory (Bytes)	Processing Time (mS)	Avg.Response Time(mS)
OAUE	84.382	84.308	83.6	83.95249	6116224	1966.6	790.6
MCM	84.508	84.431	83.833	84.13092	6124217	2063.5	887.9
ECS MINER	84.418	84.41	83.722	84.06459	6146313	2076.7	901.5
ECM-BDF	85.029	85.007	84.486	84.7456839	6100636	2005.9	831.8
CDRTREE	86.307	86.319	86.325	86.3219628	6494881	2211.6	1056.7
GEDSC	87.606	87.831	87.651	87.7408961	6434304	2189.9	1025.9

Precision measures the percentage of relevant data retrieved from the data stream and is calculated by using the formula defined as

Precision=True Positives / (True positives + False positives)

Recall measures how many of the positive data returned by a classifier and is calculated by using the formula defined as.

Recall=True Positive/(True positive + False Negative)

F1 Measure computes the balance between Precision and Recall by using the formula defined as.

F1 Measure = 2* ((Precision*Recall) / (Precision+Recall))

However the method ECM-BDF achieves high memory utilization, its accuracy rate is not up to the level of the proposed model. The method OAUE is good at processing time and response time compared to other studied algorithms, its accuracy rate is also not up to the level of the proposed model. Hence the proposed model is proven as the most reliable model, since it achieves high accuracy rate in terms of accuracy, precision, recall and F1-measure and also it is negligibly low in terms of memory utilization, and processing time than other comparative contemporary models.

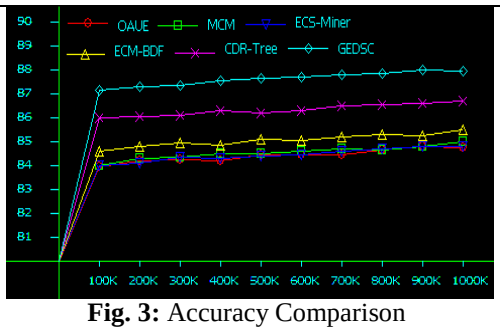

Fig. 3: Accuracy Comparison

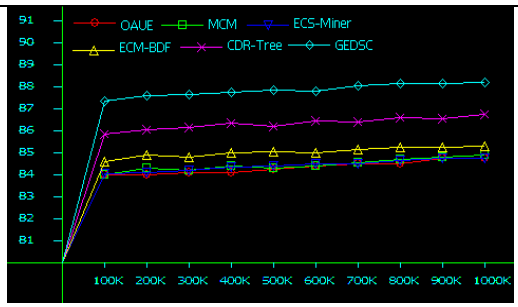

Fig. 4: Precision Comparison

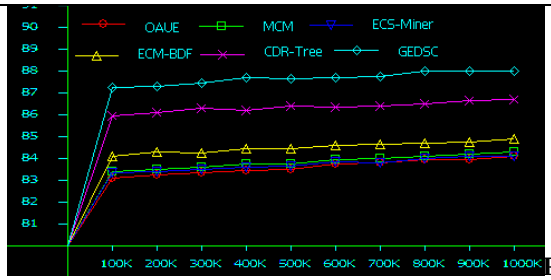

Fig. 5: Recall Comparison

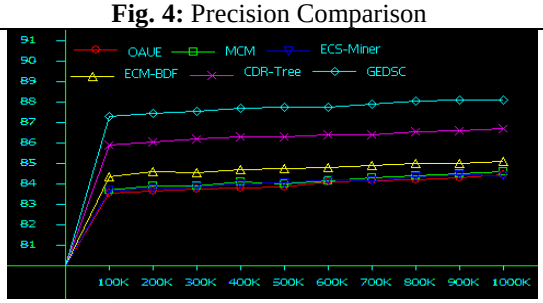

Fig. 6: F1-Measure Comparison

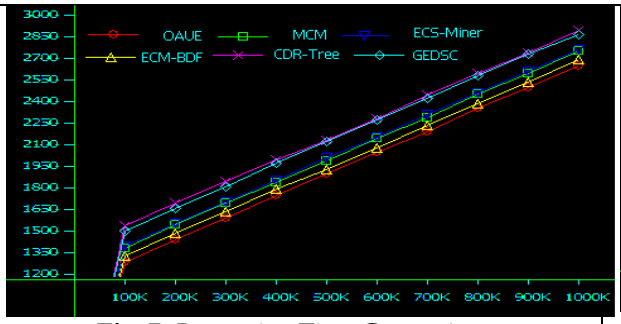

Fig. 7: Processing Time Comparison

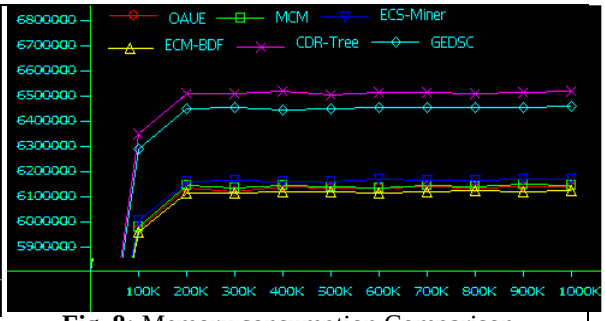

Fig. 8: Memory consumption Comparison

Conclusion and Future Enhancements:

It is observed that the proposed Genetic based Ensemble Data Stream Classifier is the most accurate of all comparative classification methods and is proved not only by calculating accuracy, but also precision, recall, and F1-measure on all varying size of data chunks.

However, its performance negligibly drops down in terms of memory utilization, processing and response time . This might be due to inherent feature of back propagation neural network algorithm which is used as a part in modified Neuro-Fuzzy data stream classifier.

Hence, it is suggested that the algorithm chosen for ensemble shall be revised to strengthen the memory utilization, processing and response time. In addition, it is also planned to deploy increased number of diverse contemporary algorithms to strengthen the classification task in all aspects. Future research is required to achieve all the above said issues and directions in data stream classification.

REFERENCES

Albert Bifet, Geoff Holmes, Bernhard Pfahringer and Ricard Gavalda, 2009. Improving Adaptive Bagging Methods for Evolving Data Streams, ACM, Proceeding ACML, pp: 23-37.

Albert Hung-Ren Ko and Robert Sabourin, 2013. Single Classifier-based Multiple Classification Scheme for weak classifiers: An experimental comparison, ACM Journal Expert Systems with Applications: An International Journal, 40(9): 3606-3622.

Charu Aggarwal, C. and Jianyong Wang, 2006. A Framework for On-Demand Classification of Evolving Data Streams, IEEE Transactions On Knowledge And Data Engineering, 18(5): 577-589.

Cheng-Jung Tsai and Wei-Pang Yang, 2008. An Efficient and Sensitive Decision Tree Approach to Mining Concept-Drifting Data Streams, Journal Informatica, 19(1): 135-156.

Cheng-Jung Tsai, Chien-I. Lee and Wei-Pang Yang, 2009. Mining decision rules on data streams in the presence of concept drifts, Expert Systems with Applications, 36: 1164-1178.

Chunquan Liang, Yang Zhang, Peng Shi and Hu. Zhengguo, 2012. Learning very fast decision tree from uncertain data streams with positive and unlabeled samples, ACM Journal Information Sciences: an International Journal, 213: 50-67.

Dariusz Brzezinski and Jerzy Stefanowski, 2014. Combining block-based and online methods in learning ensembles from concept drifting data streams, Information Sciences, 265: 50–67.

Dariusz Brzezinski and Jerzy Stefanowski, 2014. Reacting to Different Types of Concept Drift, IEEE Transactions on Neural Networks and Learning Systems, 25(1): 81-94.

Dewan Farid, Md., Li Zhang, Alamgir Hossain, Chowdhury Mofizur Rahman, Rebecca Strachan, Graham Sexton, and Keshav Dahal, 2013. An Adaptive Ensemble Classifier for Mining Concept-Drifting Data Streams, Elsevier, Expert Systems with Applications, 40(15): 5895–5906.

Domingos, P. and G. Hulten, 2000. Mining high-speed data streams, Proceedings of 6th ACM SIGKDD, pp: 71-80.

Haixun Wang, Wei Fan, S. Philip Yu and Jiawei Han, 2003. Mining concept-drifting data streams using ensemble classifiers, Proceedings of the 9th ACM SIGKDD, pp: 226–235.

Hang Yang and Simon Fong, 2013. Incremental Optimization Mechanism for Constructing a Decision Tree in Data Stream Mining, Hindawi Publishing Corporation, Mathematical Problems in Engineering, Article ID 580397.

Jayanthi, S., B. Karthikeyan, 2014. A Recap on Data Stream Classification, Adv. in Nat. Appl. Sci., 8(17): 76-82.

Jayanthi, S., B. Karthikeyan, 2014. Aggregate Weighted Ensemble Model for Data Stream Classification, International Journal of Applied Engineering Research (IJAER), 9(21), Special Issues, pp: 4945-4949.

Jayanthi, S., B. Karthikeyan, 2015. Incremental Aggregation Model For Data Stream Classification, ARPN Journal of Engineering and Applied Sciences, 10(8): 3828-3832.

Jesse Read, Albert Bifet, Bernhard Pfahringer and Geo Holmes, 2012. Batch-incremental versus instance-incremental learning in dynamic and evolving data, Proceeding IDA¨12, Springer-Verlag Berlin, pp: 313-323.

Jimenez Gonzalo Ramos, Jose del Campo-Ávila,

Rafael Morales-Bueno, 2006. Incremental Algorithm Driven by Error Margins, Proceedings 9th International Conference, pp: 358-362.

Jing, LIU., XU. Guo-sheng, Shi-hui ZHENG, Da XIAO, Gu. Li-ze, 2014. Data streams classification with ensemble model based on decision-feedback classification, The Journal of China Universities of Posts and Telecommunications, 21(1): 79–85.

Joao Gama, Raquel Sebastiao and Pedro Pereira Rodrigues, 2013. On evaluating stream learning algorithms, Springer, Machine Learning, 90(3): 317-346.

Mohammad Masud, M., 2013. Classification and Adaptive Novel Class Detection of Feature-Evolving Data Streams, IEEE Transactions On Knowledge And Data Engineering, 25(7): 1484-1497.

Mohammad, M. Masud, Latifur Khan, Jiawei Han and Bhavani Thuraisingham, 2011. Classification and Novel Class Detection in Concept-Drifting Data Streams under Time Constraints, IEEE Transactions on Knowledge And Data Engineering, 23(6): 859-874.

Wenyu Zang, Peng Zhang, Chuan zhou and Li. Guo, 2014. Comparative study between incremental and ensemble learning on data streams: Case study", Journal of Big Data, 1(5): 1-16.