iJRASET
International Journal For Research in Applied Science and Engineering Technology

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Volume: 4 **Issue:** XII **Month of publication:** December 2016

DOI:

www.ijraset.com

Call: 08813907089 | E-mail ID: ijraset@gmail.com

Volume 4 Issue XII, December 2016
ISSN: 2321-9653
International Journal for Research in Applied Science & Engineering Technology (IJRASET)

Ensemble based Classification Techniques for Concept Drifting in Continuous Data Stream: A Survey

Girish B Umaratkar[1], Jaykumar S Karniwar[2]
[1]Student, [2]Assistant Professor, Computer Science & Engineering Department
Jagadambha College of Engineering & Technology, Yavatmal, Maharashtra, India.

Abstract: ***Data Stream Mining is a process of extracting and analyzing the hidden, predictive, knowledge based information from the rapid, fast moving and raw data streams. The technical areas of data stream mining process includes Classification, Clustering, Decision Tree, Association Rule Mining, Temporal Data Mining, Time Series Analysis, Spatial Mining, Web Mining etc. From these technical areas, Stream data classification suffered from a problem of infinite length, concept evaluation, feature evaluation and concept drift. The most challenging problem of data stream is concept-drift which refers to the deviation of data stream from one state to another unpredictable state over time. For example, vital signals of human body like ECG (Electrocardiogram), EEG (Electroencephalogram), and BP (Blood Pressure) etc. are continuous in nature and abruptly changing hence there is a need to apply an efficient real-time data stream mining techniques for taking intelligent health care decisions. In order to address concept drift evolved in these continuous data stream, a classification model must endlessly adapt itself to the most recent concept. Hence, this paper gives the overview of various ensemble based classification algorithm techniques in the field of data stream mining and explores the future directions.***
Keywords: Concept Drift,data Stream Mining, Ensemble,Predictive, Rapid.

I. INTRODUCTION

Data Stream Mining is the process of extracting useful information from continuous, rapid data stream. Decision support systems usually require real-time prediction and classification based on multivariate data that have many attributes and terms. To acquire knowledge base from raw data, emphasis is placed on innovative data stream mining concepts and techniques.

While processing the data noise, errors, unwanted data, missing values have to be removed. There are many proposed classification algorithms for concept drifting data streams. A variety of techniques have also been proposed in the literature for addressing concept drift [1], [2]. These algorithms support multidimensional analysis and decision making.

Additional data analysis techniques are required for in-depth analysis, characterization of data changes over time. In addition, huge volumes of data can be accumulated beyond databases and data warehouses. In applications like video surveillance, weather forecasting, telecommunication, sensor networks, satellites, call records, vital signals monitoring; data stream mining plays a key role to analyze the continuous data. The meaningful, effective and efficient analysis of this data in such different forms becomes tedious task and also the issue of memory constraints has to handle as enormous data is generated continuously.

Figure 1, shows the general process of Data Stream Mining. The various types of data such as Internet traffic, healthcare data are taken as input and after applying data stream approaches on these inputs the meaningful, knowledge data is extracted as the result from the raw data.

Volume 4 Issue XII, December 2016
ISSN: 2321-9653
International Journal for Research in Applied Science & Engineering Technology (IJRASET)

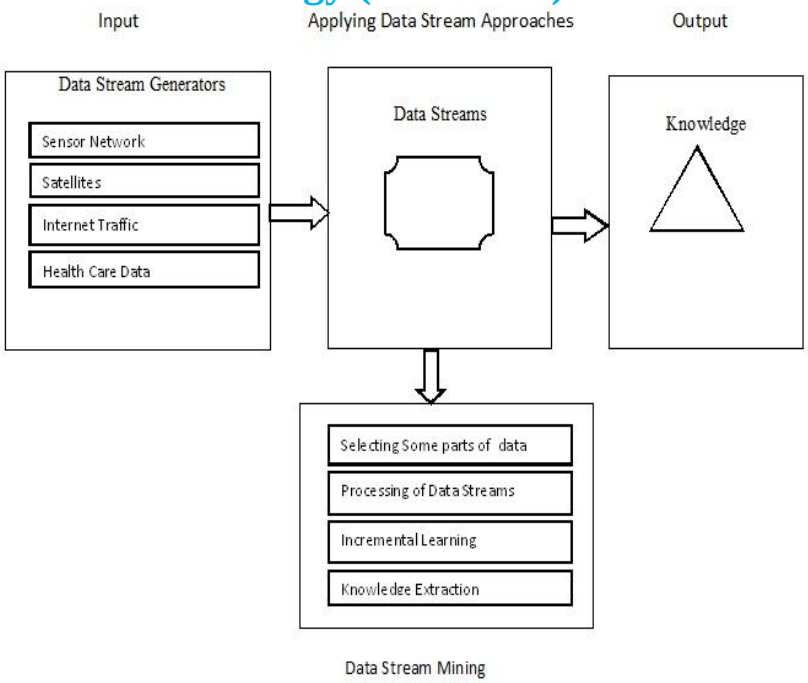


Fig. 1: General Process of Data Stream Mining [3]

II. ISSUES AND PROBLEMS IN CONCEPT DRIFT [4]

A. *Robustness*

The noise problem is more crucial for stream data mining, because it is difficult to distinguish noise from changes caused by concept drift. If an algorithm is too eager to adapt to concept changes, it may over fit noise and might be interpreting it as data from a new concept. If an algorithm is too old fashioned and slow to adapt, it may overlook important changes.

B. *Adaptation*

The concept generating a data stream drifts with time due to changes in the environment. These changes cause the model learned from old data is obsolete, and model updating is necessary.

C. *Performance*

To assure on-line responses with limited resources, continuous mining should be “fast and light”, that is:

1) Learning should be done very fast, preferably in one pass of the data.
2) Algorithms should make light demands on memory resources.

D. *Sampling data from a stream*

Value or set of values at a point in time and/or space.

E. *Filtering a data stream*

Extract only the specific data that you want to see, and then display it in the manner that you want to see it. To address these issues, analysis of distinct algorithms and strategies is required with modest resource consumption.

III. CLASSIFICATION METHODS FOR DATA STREAM

In this section, the focus is on the methods for ensemble based classification algorithm techniques in the field of data stream mining for minimization and removal a problem of concept drift. General reasons for selecting the algorithms are as follows [1].

A. *Popularity*
B. *Flexibility*
C. *Handling high dimensionality*

D. Applicability

Prof. Dipti D. Patil, Jyoti G. Mudkanna, Dnyaneshwar Rokade, Dr. Vijay M. Wadhai [3] focuses on the various ensemble based classification algorithms such as Ensemble Building, Training the Dynamical Discriminative Model, Adaptive Ensemble Classifier, Online Bagging Algorithm, DWCDS: Double Window Based Classification in real time data stream mining to calculate the efficiency and accuracy in the real-time data stream mining process for different data sets and analyzed the risk level of vital signals of human body such as ABP[dias] signal, Spo2 signal.

"Mining High Speed Data Stream" literature describes and evaluates VFDT (Very Fast Decision Tree learner), an anytime system that builds decision trees using constant memory and constant time per example. VFDT can in-corporate tens of thousands of examples per second using off-the-shelf hardware. VFDT is a decision-tree learning system based on Hoeffding trees. VFDT is I/O bound in the sense that it mines examples in less time than it takes to input from disk. It does not store any examples or parts of it in main memory, requiring only space proportional to the size of the tree and associated sufficient statistics. Hoeffding trees can be learned in constant time per example. More precisely, in time that is worst-case proportional to the number of attributes. Pedro Domingos, Geoff Hulten [4].

The paper "A Framework for On-Demand Classification of Evolving Data Stream", demonstrate that stream classification cannot be effectively performed by simply viewing it in the context of one-pass mining. The underlying classifier needs to adjust rapidly to the changes so that accurate classification results can be provided to the user on-demand. The author proposes an on-demand classification process which can dynamically select the appropriate window of past training data to build the classifier by making use of Online Algorithms for Supervised Microcluster Maintenance which discuss the process of online maintenance of microclusters and class statistics along with time. The microclusters and class statistics will be used in conjunction with a nearest-neighbor classification process in order to perform the final data stream classification. In order to perform an effective classification of the stream by using On-Demand Stream Classification Process, it is important to find the correct time-horizon which should be used for classification. In order to find the most effective horizon for classification at a given moment in time, a small portion of the training stream is not used for the creation of the microclusters. This portion of the training stream is referred to as the horizon fitting stream segment. Charu C. Aggarwal et.al. [5].

In Semi- supervised classification algorithm for data streams with REcurring concept Drifts and Limited LAbeled data called REDLLA, in which a decision tree is adopted as the classification model. When growing a tree, a clustering algorithm based on *k*-Means is installed to produce concept clusters and label unlabeled data at leaves. Potential concept drifts are distinguished and recurring concepts are maintained in the literature of Li, Xindong Wu, Xuegang Hu [6].

Ensemble algorithms are sets of single classifiers (components) whose decisions are aggregated by a voting rule. The paper "Ensemble Classifier for Drifting Concepts", Martin Scholz and Ralf Klinkenberg [7] states that the combined decision of many single classifiers is usually more accurate than that given by a single component. The author proposed Adapting Ensemble Methods to Drifting Streams which includes Ensemble Generation by KBS (Knowledge-Based Sampling) algorithm also applied a KBS-strategy to learn drifting concepts from data streams.

"Adaptive Ensemble Boosting Classifier for Concept Drifting Stream Data" literature describes the ensemble classifiers for real time data stream. This adaptive ensemble method uses boosting, adaptive sliding window and Hoeffding tree with naïve bayes adaptive as base learner for improvement of performance in data streaming and achieves distinct features as it is dynamically adaptive, uses less memory and processes data fastly by Kapil K. Wankhade and Snehlata S. Dongre [8].

Mahnoosh Kholgi, Mohhamadreza Keyvanpour [9] have been proposed the solution based on data stream mining problems and challenges. These solutions can be categorized to data-based and task-based solutions. In this classification given by author, data-based techniques refer to summarizing the whole dataset or choosing a subset of the incoming stream to be analysed and task-based techniques are those methods that modify existing techniques or invent new ones in order to address the computational challenges of data stream processing. Approximation algorithms, sliding window and algorithm output granularity represent this category.

Mohammad M. Musad et.al [10] proposes a novel technique to overcome the problem of poorly trained classifiers with the limited amount of training data, by building a classification model from a training set having both unlabelled and a small amount of labelled instances. This model is built as micro-clusters using semi-supervised clustering technique and classification is performed with κ-nearest neighbour algorithm. An ensemble of these models is used to classify the unlabelled data. Empirical evaluation on both synthetic data and real traffic reveals the approach, using only a small amount of labelled data for training, outperforms state-of-the-art stream classification algorithms that use twenty times more labelled data.

Peng Zhang et.al [11], proposed a new ensemble model which combines both classifiers and clusters together for mining data

streams. The main challenges of this new ensemble model include,

(1) Clusters formulated from data streams only carry cluster IDs, with no genuine class label information.

(2) Concept drifting underlying data streams makes it even harder to combine clusters and classifiers into one ensemble framework.

To handle challenge (1), author present a label propagation method to infer 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 (2), author present a new weighting schema to weight all base models according to 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.

According to Jing Gao, Wei Fan, Jiawei Han, [12] most existing work makes the implicit assumption that the training data and the yet-to-come testing data are always sampled from the "same distribution", and this "same distribution" evolves over time. Authors demonstrate that this may not be true, and one actually may never know either "how" or "when" the distribution changes. This paper formally and experimentally demonstrate the robustness of a model averaging and simple voting-based framework for data streams, particularly when incoming data "continuously follows significantly different" distributions. On a real streaming data, this framework reduces the expected error of baseline models by 60%, and remains the most accurate compared to those baseline models.

J. Zico Kolter, Marcus A. Maloof [13] describes an ensemble method designed expressly for tracking concept drift. Authors present an ensemble method for concept drift that dynamically creates and removes weighted experts in response to changes in performance. The method, dynamic weighted majority (DWM), uses four mechanisms to cope with concept drift: It trains online learners of the ensemble, it weights those learners based on their performance, it removes them based on their performance, and it adds new experts based on the global performance of the ensemble.

Prokhorov et.al [14] demonstrate a possibility of determining the instantaneous phases and instantaneous frequencies of the main rhythmic processes governing the cardiovascular dynamics in humans from heart rate variability data with the methods using bandpass filtration, empirical mode decomposition and wavelet transform.Human cardiovascular system (CVS) is one of the most important physiological systems whose operation is governed by several rhythmic processes interacting with each other. The analysis of the various signals such as ECG, blood pressure, blood flow and heart rate variability has revealed that they contain several almost periodic frequency components. After deriving the main rhythmic components of the cardiovascular system from the sequence of R–R intervals from such complex signal one can define their phases and examine synchronization between the rhythms.

Rajiv Kumar Nath et.al. [15] describes technique, which extracts important features from the ECG signal data in semi -automatic detection of RR interval of an ECG signal in which the user specifies the range. The RR peaks of an ECG signal are detected with of finding out the maximum value of amplitude wave in the ECG signal. This maximum value should correspond to the one pick value of the R wave in the ECG signal. After this the lower and maximum range of the R amplitude of the ECG signal is fixed with the help of data heuristics.

Dipti Durgesh Patil and Vijay M. Wadhai [16] presents the innovative wireless sensor network based Mobile Real-time Health care Monitoring (WMRHM) framework which has the capacity of giving health predictions online based on continuously monitored real time vital body signals acquired of human body such as ECG, EEG, BP, SpO2 etc. through Wireless Body Area Network (WBAN) and predict the health risk of the monitored person. The WMRHM framework is innovative as it dynamically adapts to the changes happened in the vital signals and updates the model for health risk predictions. The probability of accurate predictions is high enough as it considers Historical rule base, domain expert's rule and real-time rule model for analyzing the health status. Author also discussed Different ensemble based classifier systems such as Ensemble Building, Training the Dynamical Discriminative Model, Adaptive Ensemble Classifier, Online Bagging Algorithm, DWCDS: Double Window Based Classification in real time data stream mining algorithm for Real Time Data Stream Mining (RT-DSM). All these methods are capable of performing any-time classification, learning in one scan and detecting drift in the underlying concept. Motivation of the work is to provide mobility to patients through wireless networks and helping doctors to take preventive actions immediately by assisting them through real time decision support system.

The paper "DDD: A New Ensemble Approach For Dealing With Concept Drift" Leandro L. Minku et.al [17] presents an analysis of low and high diversity ensembles combined with different strategies to deal with concept drift and proposes a new approach Diversity for Dealing with Drifts (DDD) to handle drifts. DDD maintains ensembles with different diversity levels, exploiting the advantages of diversity to handle drifts and using information from the old concept to aid the learning of the new concept. It has better accuracy than Early Drift Detection Method (EDDM) mainly when the drifts have low severity or low speed, due to the use of ensembles with different diversity levels. DDD has also considerably good robustness to false alarms. When they occur, its accuracy

is better than EDDM's also during stable concepts due to the use of old ensembles. Besides, DDD's accuracy is almost always higher than Dynamic Weighted Majority (DWM) both during stable concept and after drifts. So, DDD is accurate both in the presence and in the absence of drifts.

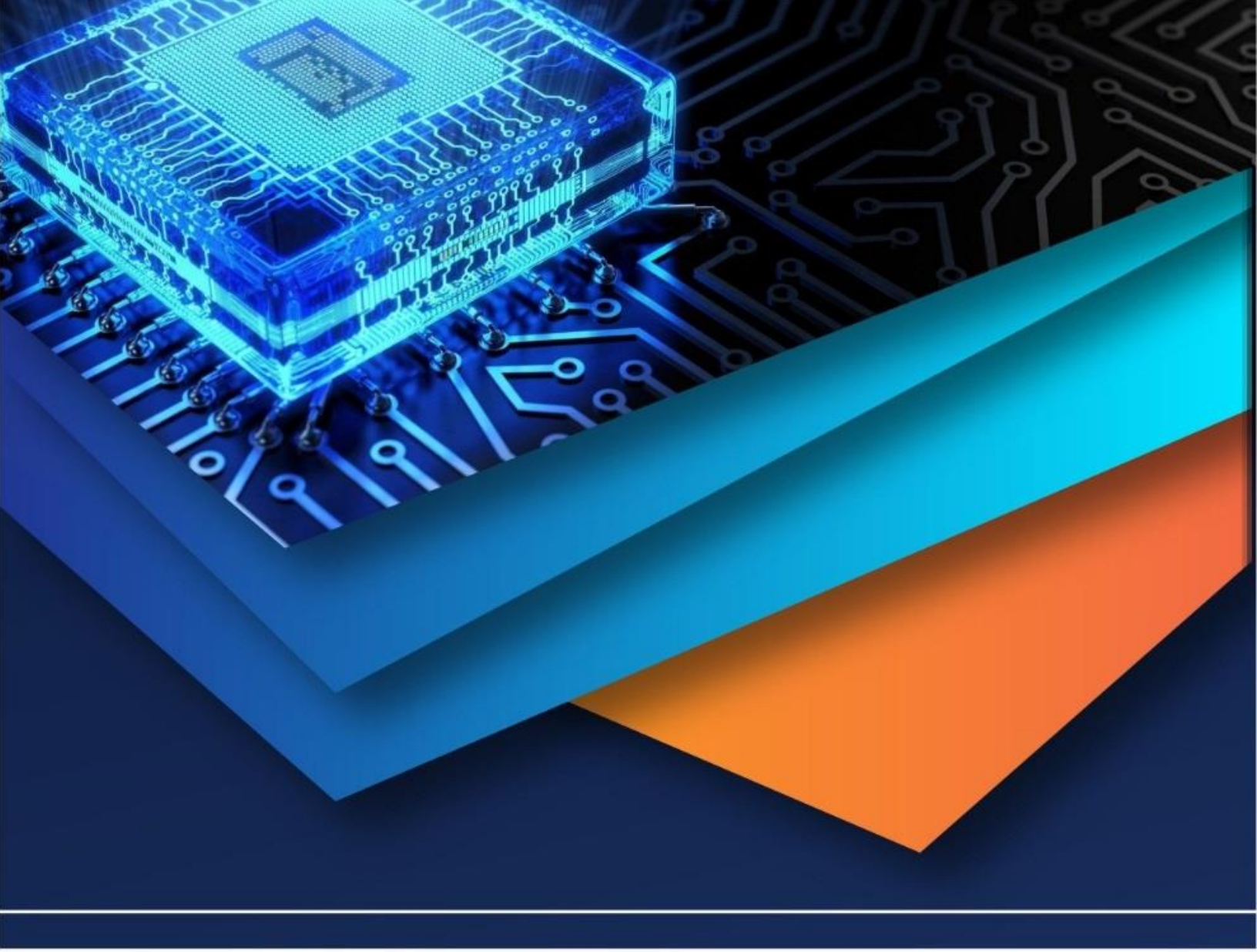
IV. CONCLUSION AND FUTURE DIRECTION

In this paper we review a various methods for ensemble based classification algorithm techniques of data stream. All these methods are capable of performing any-time classification, learning in one scan and detecting drift in the underlying concept. The Important issue of adapting concept drifts has been discussed.

The same data will be taken in real-time and dynamic ensemble based algorithms will be applied on the continuous, rapid, massive data stream to achieve maximum efficiency and accuracy in classification of data stream.

REFERENCES

[1] Yan-Nei Law and Carlo Zanily entitled, "An Adaptive Nearest Neighbor Classification Algorithm for Data Streams", in PKDD, 2005, LNAI 3721, pp. 108–120, 2005.
[2] Li Su, Hong-yan Liu, Zhen-Hui Song, "A New Classification Algorithm for Data Stream", I. J. Modern Education and Computer Science, 2011, 4, pp. 32-39
[3] Prof. Dipti D. Patil, Jyoti G. Mudkanna, Dnyaneshwar Rokade, Dr. Vijay M. Wadhai, "Concept Adapting Real-Time Data Stream Mining for Health Care Applications", Journal of Springer, ISSN: 1867-5662, Vol. 166, 2012, pp. 341-351.
[4] Pedro Domingos, Geoff Hulten,"Mining HighSpeed Data Streams", Sixth International Conference on Knowledge Discovery and Data Mining, Boston, MA: ACM Press, 2000, pp. 71-80.
[5] Charu C. Aggarwal, Jiawei Han, Jianyong Wang and Philip S. Yu, Fellow, IEEE, "A Framework for On-Demand Classification of Evolving Data Streams", IEEE Transactions on Knowledge and Data Engineering, Vol. 18, No. 5, may 2006, pp. 577-589.
[6] Peipei Li, Xindong Wu, Xuegang Hu, "Mining Recurring Concept Drifts with Limited Labeled Streaming Data" JMLR: Workshop and Conference Proceedings 13: 2nd Asian Conference on Machine Learning (ACML2010), Tokyo, Japan, Nov. 8-10, 2010, pp. 241-252.
[7] Martin Scholz and Ralf Klinkenberg,"An Ensemble Classifier for Drifting Concepts", In Intelligent Data Analysis (IDE), Special Issue on Incremental Learning Systems Capable of Dealing with Concept Drift, Vol.8, No.3, 2004, pp. 281—300.
[8] Kapil K. Wankhade and Snehlata S. Dongre, "A New Adaptive Ensemble Boosting Classifier for Concept Drifting Stream Data", International Journal of Modeling and Optimization, ISSN: 2010-3697, Vol. 2, No.4, August 2012, pp. 493-497.
[9] Mahnoosh Kholghi, Mohammadreza Keyvanpour, "An Analytical Framework for Data Stream Mining Techniques Based on Challenges and Requirements", International Journal of Engineering Science and Technology, ISSN : 0975-5462, Vol. 3, No. 3, Mar 2011, pp. 2507-2513.
[10] Mohammad M. Masud, Jing Gao, Latifur Khan, Jiawei Han, "A Practical Approach to Classify Evolving Data Streams: Training with Limited Amount of Labelle Data", IEEE International Conference on Data Mining, DOI: 10.1109/ICDM.2008.152, 2008, pp. 929-934.
[11] Peng Zhang, Xingquan Zhu, Jianlong Tan, Li Guo, "Classifier and Cluster Ensembles for Mining Concept Drifting Data Streams", IEEE International Conference on Data Mining - ICDM, DOI: 10.1109/ICDM.2010.125, 2010, pp. 1175-1180.
[12] Jing Gao, Wei Fan, Jiawei Han, "On Appropriate Assumptions to Mine Data Streams: Analysis and Practice", IEEE International Conference on Data Mining - ICDM , DOI: 10.1109/ICDM.2007.96, 2007, pp. 143-152.
[13] J. Zico Kolter, Marcus A. Maloof, "Dynamic Weighted Majority: An Ensemble Method for Drifting Concepts", Journal of Machine Learning Research: 2755-2790, Vol. 8, December 2007, pp. 2755—2790.
[14] Prokhorov M.D., Ponomarenko V.I., Gridnev V.I., Bodrov M.B., Bespyatov A.B, "Deriving main rhythms of the human cardiovascular system from the heartbeat time series and detecting their synchronization", Chaos, Solitons and Fractols (Elsevier), Vol. 23, Issue 4, February 2005, pp. 1429-1438.
[15] Rajiv Kumar Nath, Sanjay Nath,," Mining of ECG Signal for New Diagnostic Information", Indian Journal of Computer Science and Engineering, Vol. 1, No 2, 2010, pp.108-113.
[16] Dipti Durgesh Patil and Vijay M. Wadhai, "Adaptive Real Time Data Mining Methodology for Wireless Body Area Network based Healthcare Applications", Advanced Computing: An International Journal (ACIJ), Vol.3, No.4, July 2012, pp.59-70.
[17] Leandro L. Minku, Student Member, IEEE, and Xin Yao, Fellow, IEEE, "DDD: A New Ensemble Approach For Dealing With Concept Drift", IEEE Transactions on Knowledge and Data Engineering, Volume 24, Issue 4, April 2012, pp. 619-633.



10.22214/IJRASET

45.98

IMPACT FACTOR: 7.129

IMPACT FACTOR: 7.429

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY