DATA STREAM MINING : A RECAP

Dr.S.Jayanthi.
Professor,
Department of Computer Science and Engineering,
Samskruti College of Engg and Technology,
Ghatkesar – 501 301, Hyderabad, India.
nigilakash@gmail.com

Abstract - Upright growth and prevalence of communication and computing technologies have persuaded all digitized organizations and generated insurmountable streams of data in their locale. Intensive analysis of these data streams has become imminent and intrinsic need to better promote their strategic decisions and regular doings. This paper provides a brief review of the concepts required to better understand the process of data stream mining and its issues, namely, concept drift, concept evolution, outlier detection, and unboundedness. It begins with the discussion of data mining and the role of the classification task in data analysis. Then it deliberates about data stream mining and data stream classification task. It also outlines the deficits of conventional classification approaches on achieving data streaming environment and the need for promoting research on data stream classification.

Keywords: Data streams, Data stream classification, concept drift, concept evolution, data analysis

I. INTRODUCTION

Breakthrough in computing technologies has facilitated the generation of a huge volume of data in the industrial locale. Data mining is a prominent domain fertilized to satisfy the unsolvable demands of database technology which has been used prior to the inception of data mining to process the data stored in the databases (Bifet *et al*., 2009; Kosina *et al*., 2012).

Data mining is concerned to extract previously unknown and hidden knowledge in massive data repositories. It is an interdisciplinary domain which imports conventions and principles suitably from machine learning, statistical learning, pattern recognition, database technology, intelligent systems, artificial intelligence, visualization technology and other disciplines. In recent years, streams of data are floated over e-industries. Analyzing these infinite data streams has become a highly challenging task, as they are generated from diverse medium and in diverse forms at a faster rate than ever before. Data streams are massive, dynamic and infinite in nature and arriving from diverse dynamic distribution centres. Hence, processing data streams in a resource aware environment poses several challenges (Gama *et al*., 2013).

In general, data streams are expected to be processed using a single pass scan to comply with constrained resource usage in an online environment. These constraints have imposed several obstacles on conventional data mining algorithms which are only efficient in processing data stored in static storage medium, with multi pass scanning, where the data stored are bounded and predictable (Gomes *et al*., 2014; Masud *et al*., 2009)

In light of addressing the deficits of data mining algorithms, data stream mining has been stemmed from data mining and turned into an active research spot. Data stream mining is intended to process massive data streams in a resource constrained environment. The research in data stream mining is concerned on empowering data stream mining process either by exploring fine-tuned version of available data mining algorithms or distinct novel data stream algorithms.

II. Knowledge Discovery in Databases (KDD) Process

In general, the KDD process refers to the overall process of discovering useful and interesting patterns of the stored data in large repositories. KDD process is focused on developing tools to control the flood of data floated over e-industries. Knowledge discovery process in a database employs five distinct operations, namely, selection of task relevant data set, preprocessing, transformation, data mining, and evaluation or interpretation results (Han *et al*., 2006).

This process starts with extracting task relevant data from the large repository of data by using data preprocessing techniques such as data cleaning, selection, integration, and transformation. The resultant cleansed and compact data from preprocessing techniques are transformed into the format suitable for data analysis. Then suitable data

mining tasks are applied to analyze the transformed data to produce interesting patterns. Further, the produced patterns undergo for evaluation so as to assess the quality of it. Finally, patterns are interpreted using efficient visualization techniques. However, data mining constitutes as a subtask in KDD process, it is evident from the increasing demands of analyzing massive data makes data mining as the most inevitable task and has captivated the attention of many researchers over a couple of decades.

A. Data Mining Tasks

Despite the availability of a range of tasks in data mining process, they can be classified into two major categories, namely, predictive mining tasks and descriptive mining tasks.

Predictive mining tasks perform pertinent analysis by making inferences on the stored data. On the contrary, descriptive mining task is intended to make valuable insight on the stored data by describing its typical general characteristics. Data mining tasks shall be suitably chosen based on the need of application in which it is deployed (Han and Kamber; 2006)

Classification

Classification task constructs a classification model which performs supervised learning using training data and classification rules to acutely classify the testing data into predetermined class labels. Classification model selects a suitable classification algorithm, such as decision tree, neural networks, naïve Bayesian, etc., to achieve the given classification task. However each classification algorithm holds its own strength and weakness pertaining to the nature of data set and application.

Regression Analysis

The regression analysis task is intended to develop a model that uses existing values to forecast the underlying trend or the associated values in the unknown data using statistical methods.

Clustering

Clustering process undertakes unsupervised learning to group the data into several classes or clusters which are not predetermined, by maintaining high intra class similarity and inter class dissimilarity measure between the clusters.

Relevance or Association Analysis

Relevance or Association analysis finds out an interesting relationship between values of variables by formulating association rules along with interestingness measures, namely, support and confidence. If interestingness measures are less than the specified threshold values, they can be discarded as uninteresting.

Summarization

Summarization provides a compact representation of data set by using visualization and report generation techniques.

Among the above discussed data mining tasks, classification is the most inevitable and widely used task for analyzing data (Tsai *et al*., 2007). It has been observed that many research works are attracted to investigate the classification task.

III. BASIC PRINCIPLES OF CLASSIFICATION APPROACHES

In general, classification approaches are supervised in nature, where the model created for classification is trained to classify the preset number of classes with preset classification rules.

A. Classification Model

Classification model of a classifier is designed with a right choice of classification algorithms and classification rules. Each classifier employs two phases, namely, training phase and testing phase. In the training phase, the model is trained to learn the training data with predetermined class labels and preset classification rules for each class label. In the testing phase, the classification model is deployed to scan and classify the testing data based on the learning gained in the training phase and the efficacy of the classifier is ensured by comparing its accuracy with the preset threshold value of accuracy.

The test data can be either chosen from training data set, where the nature of dataset is known, or general data set where the nature of data is unknown and not used for training the classification model. The classifiers accuracy is evaluated by calculating the percentage of the number of correctly classified tuples in the test data. The classification process is depicted in Fig.1.

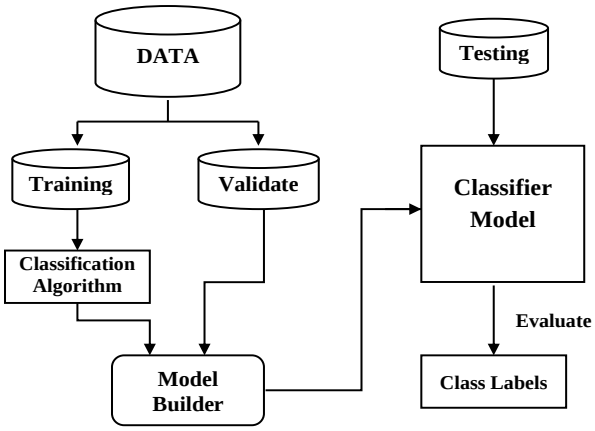

Fig.1. Classification process

B. Classifier Accuracy Estimation Techniques

Accuracy estimation techniques are used to evaluate the correctness of the classifier in testing phase of the classification process. Holdout, bootstrap and cross validation are widely used for evaluating the accuracy of the classifier (Han and Kamber, 2006; Japkowicz and Shah, 2011)

Holdout Method

It evaluates the accuracy of the classifier by segmenting the two third of the data set for training and the remaining portion for assessing the accuracy of a classification task. Herein, the efficacy of the evaluation is subject to the training data set of the classification task.

Random subsampling method is a variant of it which determines the accuracy of the classification result by calculating the percentage of the accuracy obtained by the repeated iterations of the holdout operation.

Cross validation

It evaluates the accuracy of the classifier by repeating the holdout method for k times on k segments of the given data set where each data becomes training dataset for k-1 times and testing dataset for exactly once.

Bootstrap Validation

It evaluates the classifier accuracy with random sampling replacement where each selected training tuple is equally likely to become training tuple again in the future.

Deficits of Conventional Classification Approaches on Data Stream Classification

The technical deficiency of conventional classification approaches on analyzing data streams is delineated underneath (Brzezinski *et al*., 2014):

Data mining algorithms

- are trained in static environments where resources and training data are bounded and fixed
- wait until the complete arrival of the dataset from a distribution centre
- perform multiple scanning of data by storing the entire dataset in a stable storage medium

The above stated features of data mining algorithms lead to flaw in achieving data stream mining task where single scanning, instant response, and constrained resource usage are the foremost demands.

IV. DATA STREAM MINING

Leading to the shortfalls of data mining algorithms on analyzing data streams in a real time mode, data stream mining has been explored. The process of discovering knowledge and information from unbounded and fast evolving data streams is known as data stream mining (Yang *et al*., 2011, Gaber *et al*., 2005).

Distinguished features of data streams include:

- Ever growing volume
- Infinite length
- Fast evolving
- Rapid and temporal sequence

Data Stream Pre-Processing Techniques

Data stream pre-processing techniques are inevitable in data analysis as they produce approximate answers in more compact form. These techniques can broadly be categorized into two major types, namely, data based processing and task based processing

Data Based Processing Techniques

Data based processing or summarization techniques generally provide approximate solutions either by scanning the whole dataset or the subset of data set.

Task Based Processing Techniques

Task-based techniques modify the existing data stream classification techniques and by which formulate new classifier in order to meet with the computational challenges of data stream processing.

Window based techniques, approximation approaches, and Algorithm Output Granularity (AOG) are the most popularly used tasks based processing techniques.

V. DATA STREAM CLASSIFICATION

Among several tasks of data stream mining, data stream classification is the most crucial task, widely sought to carry out online analysis. Data stream classification methods predict and classify the testing data streams based on the obtained learning and experience from training data and application domain through preset classification rules under constrained resources (Kuncheva *et al*., 2008; Bifet *et al*., 2009).

Conventional classification techniques perform block based learning on small datasets here the whole training data is available to the learning alg orithm and data instances are processed multiple times with the assumption that the instances are generated by a stationary distribution centre. But, batch learning approaches are not successful when applied to highly fluctuating data streaming environment (Gomes *et al*., 2014).

Most popular categories of data stream classification approach visited in the literature are (Tsai, 2007):

- Window Based Approaches
- Weight/Aging based Approaches
- Ensemble Learning Approaches
- Incremental Learning Approaches

A. Window Based Approaches

Window based Approaches are more suitable for applications which are interested only in analyzing the recent history of the data streams (Tsymbal *et al*., 2008). Among several window based approaches, fixed sliding window model, Adaptive Window model (ADWIN), landmark window model, Probabilistic Approximate Window model (PAW) and Self-Adaptive Sliding Window model (SASW) are more popular.

Handling of the unbounded length of data streams is also a highly notable concern of data stream classifiers which are practiced with a range of techniques to cope with this issue. However, sliding window approaches have been widely adopted to cope with the unbounded length of data stream classifiers. Sliding window model has been widely adopted to carry out pre-processing as well as to confront the constraints of the data stream classification process.

B. Weight Based Approaches/Aging techniques

In this approach, weights are assigned to the instances of data streams with respect to its age, which is calculated based on its arrival time into the window in which the instances will be scanned. Damped window model and a family of uniFied Instance Selection algoritHm (FISH algorithms) are the most widely used weight based learning approaches.

Damped window model upholds a window with a decay function that assigns a weight to each instance of data streams on its arrival, and lowers its weight exponentially over the time. Fixed Sliding Window and Damped Window will not make an analysis on the historical data while others do so.

FISH algorithms (uniFied Instance Selection algoritHm) use time and space information to establish a window for training instances at each time step. FISH1 algorithm which is the revised version of FISH algorithms uses pre specified fixed sized window of training instances.

C. Incremental Learning Approaches

Incremental learning based classification algorithms are efficient in learning from dynamic data streams by incrementally revising the classification model either by using the single classifier approach or the ensemble classifier approach (Farid *et al*., 2013; Jing *et al*., 2014). These approaches require less or no access to the instances of outdated data while preserving the knowledge about historical data. Incremental learning approaches also have the ability to learn novel classes.

Incremental learning algorithms perform classification by scanning the input instances one-by-one sequentially and update the model only after receiving complete instance (Read *et al*., 2012). It is adopted to make inferences on dynamic data streaming environment where the delay in classification results is tolerable.

To describe formally, a sequence of pairs of instances $(x,a)_1,(x,a)_2,...,(x,a)_t,...$, are solved by online classification algorithms, here x is the feature vector of instances and a is its class label. The class label of each instance can take any value from a finite set of decision classes $A=\{al:_{l=1,...,K}\}$, which has cardinality K. Each sequence pairs of instances are feed into learning algorithm as training instances TR.

At each step t of the training period, the classification model where the learning algorithm deployed is used to find out the best possible approximation f' of the unknown function f, where f(x)=a. Subsequently, f' can be used to find the class a= f' (x) for any x such that (x,a)∉TR. After the classification, the classification model receives the feedback on the actual class label. If the classification or prediction is not correct, then learning algorithm takes up additional next steps f't+1 using f't and (x,a)t (Czarnowski *et al*., 2014).

In single incremental learning approach, the single classifier is taught to learn with a specific learning method and tailored gracefully handle to concept drifts and novel classes. At the incidence of concept drift, this process performs complex operations to update the internal structure of the classifier, which debases the classifier's performance on the data streaming environment.

In contradictory, ensemble classifier approach contains several classifiers which are prespecified and fixed during the learning period, and should be updated at each time when concept drift occurs on its deployment in data streaming environment (Kolter *et al*., 2007; Hung et., 2013)

Incremental learning can be classified into two categories, namely, instance-incremental and batch-incremental model (Zhang *et al*., 2011). Instance-incremental approach updates the model at each time when new training instances arrive. Naive Bayes, neural networks, hoeffding decision trees, k-nearest neighbour, etc. are instance based incremental in nature.

Batch-incremental model updates the model only when n training examples available, where n is the predetermined size of the batch. Logistic regression, support vector machines, decision trees, etc, perform a batch based increment over data stream processing.

D. Single Incremental Learning Approach

Decision tree algorithms are efficient and widely adopted to perform classification in a static environment where all training data sets are stored in the main memory. Since it limits the size of the data sets with respect to that of the storage medium, it is not suitable for learning data streams. A number of decision tree based data stream learners are available to perform the data stream classification process.

Very Fast decision tree (VFDT), Concept drifting VFDT (CVFDT), VFDT for continuous attributes (VFDTc) , *Ultra Fast* Forest of Trees (*UFFT*), etc., are decision tree based data stream classifiers. The main drawback of decision tree based data stream classifiers is that they are more sensitive to concept drift by which its accuracy and computational efficiency are degraded.

VFDT (Pedro Domingos *et al*., 2000) is one of the first algorithm based on the hoeffding tree which implements hoeffding bound to choose the best splitting attribute. It is capable of learning from high speed data streams with very small constant time per instance. Its drawback is that it is efficient only in concept drift free data streaming environment where no causes of concept drift may arise. VFDT is a pioneer for the emergence of many algorithms like CVFDT, VFDTc, UFFT, etc.

CVFDT (Geoff Hulten *et al*., 2001) uses fixed sliding-window approach in addition to the principles of VFDT to deal with concept-drifts while maintaining similar efficiency and speed of VFDT. Its drawback is that it forgets out dated old data and treats recent data as more accurate and vital.

UFFT (Joao Gama *et al*., 2004) resolves the multiclass problem by generating the binary trees for each possible pair of classes. It performs the splitting test in each leaf node by using the hoeffding bound and detects concept drift by using naive bayes classifier at inner nodes and leaves. It also uses short term memory window to store the statistics of each leaf.

VFDT for continuous attributes (VFDTc) (Gama *et al*., 2006) has the ability to process numerical data and provides more efficient classification on continuous data. It also applies naive bayes classifier at binary tree leaves, but it detects concept drift by continuously monitoring the differences between class distributions.

Similar to the decision tree based data stream classifiers, a range of research work has been probed on ruled based classifiers, support vector machine based classifiers, genetic algorithm classifiers, similarity based classifiers, etc. The following are the example research works which adopt the principles of rule based classifier and similarity based classifier approach.

Aggarwal *et al*. (2006), proposed an on-demand classification technique that applies horizon fitting tactic to perform dynamic classification by adopting micro clustering approach and geometric time window

approach that helps to track up-to-date snapshots with less storage and processing overhead.

E. Online Learning Algorithms

To learn from the dynamic environment, online learning algorithms are less restrictive than incremental algorithms. It has the ability to learn from data streams even when not having the complete training data at the beginning. Here the classification model needs to be continuously updated during the arrival of fast evolving massive data streams.

As instant response is the major constraint, online learning algorithms are more preferable than incremental learning algorithms in almost all real time applications where the timely discovery of approximate knowledge is highly valuable, instead delayed accurate knowledge (Brzezinski *et al*., 2014; Masud *et al*., 2013; Minku *et al*., 2010). However, incremental learning and online learning algorithms are used alternatively in literature.

From the above stated state of the art of incremental or online data stream classification approaches, it is apparent that they all incur time and space overhead as they are practiced to learn from a single portion of data streams and they need to be updated each time upon the arrival of new instances.

Ensemble classification approaches have been frequently adopted to handle concept drifts in data stream classification as it is more prominent for high accuracy, scalability, and robustness.

F. Ensemble Learning Classifier

It is well known in the data analysis community that no single algorithm achieves high accuracy for all situations. That is, an algorithm that works well on one or more datasets may work badly on others. To address this issue, an ensemble of classifiers has been used to produce better classification results.

In ensemble classification approach, multiple classifiers are combined to solve a particular problem (Bifet, 2009, Farid *et al*., 2013) Ensemble classifier is shown in Fig.2.

Ensemble classification methods studied in the literature differ over the following dimensions:

- Choice of base classifier
- Handling of input training data
- Aggregation approach to integrate the outputs of member classifiers
- Methods used to initialize and adapt weights of the ensemble
- Retraining methods

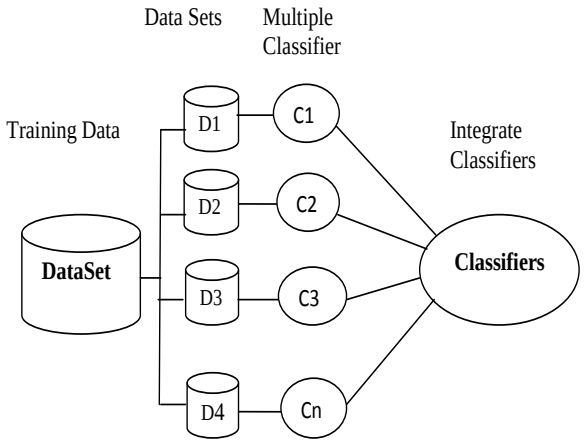

Fig.2. Ensemble classifier

Ensemble based classification approaches are preferred because of its ability to scale up the inductive algorithms with respect to the size of large databases, learn from distributed datasets, and learn from concept-drifting data streams. These abilities have attracted the attention of data stream mining researchers in investigating ensemble methods to achieve high accuracy, reliability, and robustness in data streams mining task (Tsymbal, 2008).

Ensemble classification methods are generally categorized into homogeneous and heterogeneous ensemble methods. In homogeneous ensemble approach, each classifier in the ensemble is of the same type, but each differs with respect to its' attribute list, training set, and distribution centre. In heterogeneous ensemble approach, each classifier in the ensemble is different and maintains high diversity.

Ensemble classifiers can be further divided into block-based and online approaches. Block-based ensemble approaches learn only when a block of n instances is available. It evaluates its components periodically by using weighting approaches. Online ensemble approaches learn instances immediately upon its arrival.

VI. CONCLUSION

Significant facets of the intensive study over the state of the art of data stream mining are summarized at the lower place:

- Over a span of decades, Data mining has been applied as a panacea to process data stored in massive data sets by resolving the obstacles confronted by database technologies.
- However, in recent years, streams of data have been generated in e-industries, which is impossible to make an analysis on data streams by storing it on a stable storage medium.
- Data mining algorithms which are trained to analyze the data stored in a static storage medium using multiple scanning become unsuitable as the instant response with constrained resources has become the central concern of online analysis.
- Data stream mining has emerged with several efficient data streaming algorithms to resolve this issue.
- A range of pre-processing approaches is available in the literature which can be suitably adopted before employing data streaming algorithms on data streams.
- Among several tasks of data stream mining, data stream classification has frequently been used in e-industries. Data stream classification algorithms can broadly be categorized into four approaches, namely, window based approaches, weight/aging based approaches, ensemble learning approaches and incremental learning approaches.
- Despite the availability of a wide range of approaches in the literature, data stream classification is yet in infancy stage where each approach has its own weakness regardless of its strength.
- The contexts discussed in this paper intensively emphasize the prominence and eminence of data stream mining, data stream classification process and the need for promoting it.

REFERENCES

1. Albert Bifet, Geoff Holmes, Bernhard Pfahringer and Ricard Gavalda, "Improving Adaptive Bagging Methods for Evolving Data Streams", ACM, Proceeding ACML, 2009, pp. 23-37.
2. Albert Hung, Ren Ko, and Robert Sabourin "Single Classifier Based Multiple Classifications", Proceedings, 2013, pp. 134-145.
3. *Alexey* Tsymbal, Mykola Pechenizkiy, Pádraig Cunningham and Seppo Puuronen, "Dynamic integration of classifiers for handling concept drift", Elsevier, Information Fusion, 9(1): 2008, pp.56–68.
4. Charu C. Aggarwal and Jianyong Wang, "A Framework for On-Demand Classification of Evolving Data Streams", IEEE Transactions On Knowledge And Data Engineering, 18(5): 2006, pp.577-589.
5. *Chen Li*, Yang Zhang and Xue *Li, "OcVFDT:* one-class very fast decision tree for one-class classification of data streams", Proc. SensorKDD, *2009*. pp. 79-86.
6. Cheng-Jung Tsai, Chien-I Lee, and Wei-Pang Yang "An Efficient and Sensitive Decision Tree Approach to Mining Concept-Drifting Data Streams", 2008, Informatica, 19(1): 135–156.
7. Brzezinski and Stefanowski, "Reacting to Different Types of Concept Drift: The Accuracy Updated Ensemble Algorithm", IEEE Transactions on Neural Networks and Learning Systems, 25(1):2014, pp.81-94.
8. Md. Farid, Li Zhang, Alamgir Hossain, Chowdhury Mofizur Rahman, Rebecca Strachan, Graham Sexton and Keshav Dahal, "An Adaptive Ensemble Classifier for Mining Concept-Drifting Data Streams", Elsevier, Expert Systems with Applications, 40(15): 2013, 5895–5906.
9. Hang Yang and Simon Fong, "Moderated VFDT in Stream Mining Using Adaptive Tie Threshold and Incremental Pruning", Proceeding DaWaK, 2011, pp. 371-483.
10. Jeremy Z. Kolter and Marcus A. Maloof, "Dynamic Weighted Majority: An Ensemble Method for Drifting Concepts", Journal of Machine Learning Research, 8, 2007, pp.2755–2790.
11. Jesse Read, Albert Bifet, Bernhard Pfahringer and Geo Holmes, "Batch-incremental versus instance-incremental learning in dynamic and evolving data", Proceeding IDA, Springer-Verlag Berlin, 2012, pp.313-323.
12. Jiawei Han and Kamber, "Data Mining Concepts and Techniques", Margan Kaufman Publisher, Second Edition, 2006.
13. Joao Bártolo Gomes, Mohamed Medhat Gaber, Pedro A. C. Sousa and Ernestina Menasalvas, "Mining Recurring Concepts in a Dynamic Feature Space, IEEE Transactions on Neural Networks and Learning Systems", 25(1): 2014, pp. 95-110.
14. Joao Gama, Sebastiao and Pereira Rodrigues, "Issues in Evaluation of Stream Learning Algorithms", KDD Proc. of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining, 2009, pp.329-338.
15. Joao Gama, Sebastiao and Pereira Rodrigues, "On evaluating stream learning algorithms, Machine Learning", 90(3), 2013, pp.317-346.
16. Leandro L, Minku and Xin Yao, "DDD: a new ensemble approach for dealing with concept drift", IEEE Transactions on Knowledge and Data Engineering", 24(4): 2012, pp.619-633.
17. Liu Jing, Xu Guo-sheng, Zheng Shi-hui, Xiao Da and Gu Li-ze, "Data streams classification with ensemble model based on decision-feedback", The Journal of China Universities of Posts and Telecommunications, 21(1), 2014, pp.79–85.

18. Ludmila I. Kuncheva, "Classifier ensembles for detecting concept change in streaming data: overview and perspectives", in: Proc.2nd Workshop SUEMA 2008.
19. Mohamed Medhat Gaber, Arkady Zaslavsky and Shonali Krishnaswamy, "Mining Data Streams: A Review", SIGMOD Record, 34(2), 2005, pp. 18-26.
20. Mohammad M. Masud, Jing Gao, Latifur Khan, Jiawei Han and Bhavani Thuraisingham, "Integrating Novel Class Detection with Classification for Concept-Drifting Data Streams", *ECML PKDD*, Springer-Verlag Berlin Heidelberg, 2009, pp.79–94.
21. Mohammad M. Masud, Qing Chen, Latifur Khan, Charu C. Aggarwal, Jing Gao, Jiawei Han, Ashok Srivastava and Nikunj C. Oza, "Classification and Adaptive Novel Class Detection of Feature-Evolving Data Streams", IEEE Transactions On Knowledge And Data Engineering, 25(7), 2013, pp.1484-1497.
22. Nathalie Japkowicz and Mohak Shah, "Evaluating Learning Algorithms: A Classification Perspective", Cambridge University Press, 2011.
23. *P. Domingos* and *G. Hulten., 2000. Mining high-speed data streams, Proc.* of the 6th ACM SIGKDD, 71-80.
24. Petr Kosina and Joao Gama, "Handling Time Changing Data with Adaptive Very Fast Decision Rules", Proceeding ECML PKDD, 2012, pp.827-842.

Dr.S.Jayanthi was born in the year 1981. She received her Bachelor degree in Computer Science from Bharathidasan University, India in 2002, and Master degrees in Computer Applications, and Computer Science and Engineering from Bharathidasan University, in 2005, and from Anna University of Technology, India in 2009, respectively. She obtained her Ph.D from Karpagam University, Coimbatore, India. She has 8 years of teaching experience. Her area of research includes Data Mining, Neural Networks and Big Data analysis. So far she has published 26 papers in various national, International conferences and journals.