iJRASET
International Journal For Research in Applied Science and Engineering Technology

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Volume: 5 **Issue:** VI **Month of publication:** June 2017

DOI:

www.ijraset.com

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

Volume 5 Issue VI, June 2017
ISSN: 2321-9653
International Journal for Research in Applied Science & Engineering Technology (IJRASET)

Data Stream Classification Techniques to Find Novel Class – A Review

Aswathi R Kaimal[1], Joshy Thomas Jose[2]
[1,2] Computer Science and Engineering, Kerala Technological University

Abstract*: Data Stream Mining (DSM) is the way of retrieving useful knowledge from fast information records. Today, there are number of application that produce massive amount of stream data. Concept evolution, the phenomenon of class disappearance and emergence is a valuable research topic in data stream mining field. Concept evolution happens when a new class occurs in the stream. Similar to data mining, data stream mining includes classification and clustering techniques; the special focus of this paper is on classification techniques developed to handle data streams. The problem of classification is one of the most widely studied in the context of data stream mining. The problem of classification is made more difficult by the evolution of the underlying data stream. Based on the efficiency of the classification technique used, the accuracy of classification varies. This paper reviews different classification techniques for handling the novel class in the data streams.
Keywords*: Data stream mining, classification, concept evolution, ensemble model, reliability

I. INTRODUCTION

Classification problems [1] have been studied thoroughly as a major category of the data analysis tasks in data stream mining. Data Stream Mining is the process of extracting knowledge structures from continuous, rapid data records. A data stream is an ordered sequence of instances that in many applications of data stream mining can be read only once or a small number of times using limited computing and storage capabilities. Examples of data streams include computer network traffic, phone conversations, ATM transactions, web searches, and sensor data. Data stream mining can be considered a subfield of data mining. In many data stream mining applications, the goal is to predict the class or value of new instances in the data stream given some knowledge about the class membership or values of previous instances in the data stream. In many applications, especially operating within non-stationary environments, the distribution underlying the instances or the rules underlying their labeling may change over time, i.e. the goal of the prediction, the class to be predicted or the target value to be predicted, may change over time. This problem is referred to as concept drift [3].

Data mining process has two major tasks: classification and clustering.[1]. In data stream classification supposed that total no of classes are fixed. Its not true for real environment when new classes may involve. The goal of data mining classifiers is predict the class value whose attributes value are known but class value is unknown. Classification maps data into predefined that is referred to a supervised learning because classes are determined before examining data. In clustering class or groups are not predefined but rather defined by the data alone. It is referred as unsupervised learning. Data stream classifiers may either be single model incremental approaches, or ensemble techniques, in which the classification output is a function of the predictions of different classifiers. Ensemble techniques have been more popular than their single model counterparts because of their simpler implementation and higher efficiency [1].

More importantly, the characteristics of the data stream can change over time and the evolving pattern needs to be captured. Due to the large volume and the high speed of streaming data, mining algorithms must cope with the effects of system overload. Thus, how to achieve optimum results under various resource constraints becomes a challenging task. This paper, reviews different classification techniques for data analysis. Some general issues in stream data mining are discussed.

II. DATA STREAM MINING

Mining data streams is concerned with extracting knowledge structures represented in models and patterns in non-stopping streams of information. The research in data stream mining has gained a high attraction due to the importance of its applications and the increasing generation of streaming information. Applications of data stream analysis can vary from critical scientific and astronomical applications to important business and financial ones. High volume and potential infinite data streams are generated by so many resources such as real-time surveillance systems, communication networks, Internet traffic, on-line transactions in the financial market or retail industry, electric power grids, industry production processes, scientific and engineering experiments, remote sensors

Volume 5 Issue VI, June 2017
ISSN: 2321-9653

and other dynamic environments. In data stream model, data items can be relational tuples like network measurements and call records. In comparison with traditional data sets, data stream flows continuously in systems with varying update rate. Data streams are continuous, temporally ordered, fast changing, massive and potentially infinite [4]. Due to huge amount and high storage cost, it is impossible to store an entire data streams or to scan through it multiple times. So it makes so many challenges in storage. Because of high volume of input data, it is needed to use semi-automatic interactional techniques to extract embedded knowledge from data. The data mining technique is shown in the fig 1. The collected data is given to the pre-processing step. Data preprocessing is an important step in the data mining process. Data preprocessing contains cleaning, instance selection, normalization, transformation, feature extraction and selection etc. So in this step, the irrelevant data are removed which in turn reduces the processing time. After the pre-processing step, data mining techniques are applied. Finally useful knowledge is produced.

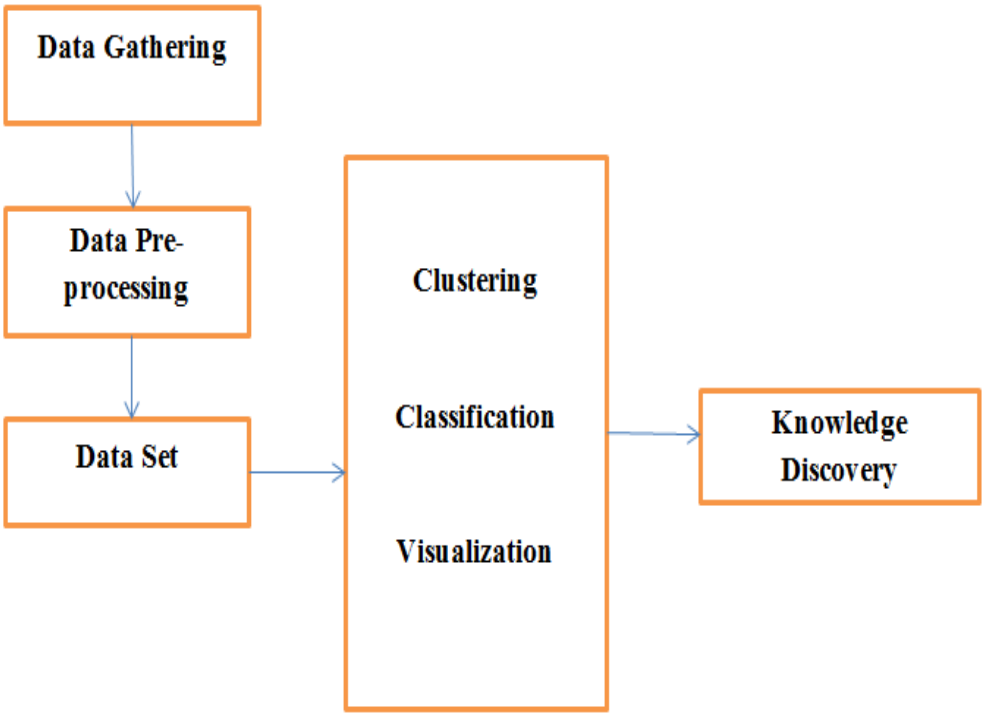


Fig. 1 Data Mining Technique

III. CONCEPT EVOLUTION

The major challenge in classifying data streams is emergence of a totally new class, also known as concept evolution. Concept evolution [2] means the emergence or disappearance of a class in the data stream. The number of classes is not fixed, and concept evolution may occur at any time in the stream. If concept evolution is not addressed timely, classifier might misclassify the instances from the new class as existing classes, thereby the classification error increases. However, this problem has been ignored by most of state-of the-art techniques. For example, consider the problem of intrusion detection in a network traffic stream. Here each type of attack is considered as a class label, and then concept-evolution occurs when a new type of attack occurs in the traffic. The problem of concept-evolution is addressed in only a very limited way by the currently available data stream classification techniques. So, the classification model needs to be updated continuously to adapt to the most recent concept.

IV. CLASSIFICATION AND NOVEL CLASS DETECTION TECHNIQUES

Classification [2] is a data mining technique used to predict group membership for data instances. Classification raises many challenges, some of which have not been addressed yet. Most existing data stream classification algorithms address two major problems related to data streams: their “infinite length”, and “concept drift”. The data stream mining process is shown in fig 2. The training texts are selected from the given data. The training text together forms a training module. This training module is

Volume 5 Issue VI, June 2017
ISSN: 2321-9653

considered as a classification model. It uses some classification rules for finding the features of the data. Then the text to be classified is given to the classification module. After the classification, the result is compared with the training data set.

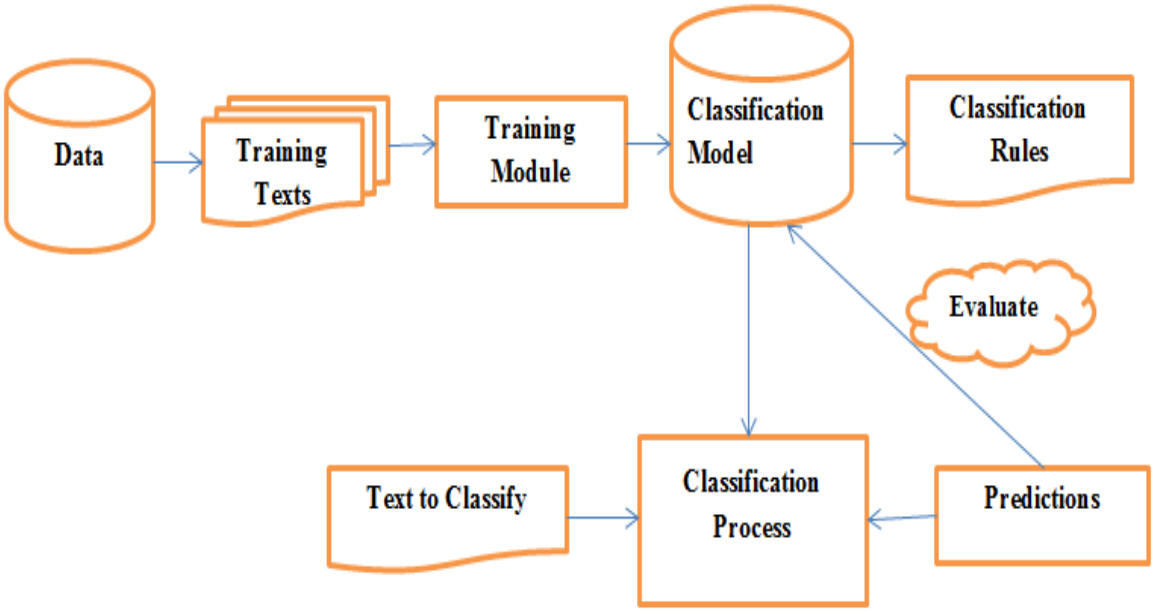


Fig. 2 Data Stream Mining Process

A. Different Techniques for Detect Novel Class

1) Cluster Based Technique: The cluster based technique [5] builds a simple model of the data using clustering technique defined by the hypersphere encompassing all the clusters of normal data. This model is continuously updated with stream evolution. If any cluster is formed outside this hypersphere, which satisfies a certain density constraint, then it is considered as a novel class. However, this approach assumes only one "normal" class, and considers all other classes as "novel". Therefore, it is not appropriate for multi-class data stream classification, since it corresponds to a "one-class" classifier.

2) Decision Tree: A new decision tree learning approach [6] for novel class detection. In this builds a decision tree from data stream which continuously update. When new a data point is arrived, it is stored in a corresponding leaf based on the values of its features. When the number of points in a leaf exceeds a predefined threshold value, the leaf points are clustered using k-means clustering method. These clusters are used to generate concepts and to identify unlabeled data using majority class classification method. Concept drift is detected if the distance between the centroids of the new data cluster and reference cluster is greater than the sum of their radius. The decision tree classifier is a data stream classification technique which builds a decision tree from data.

3) ECS Miner: ECS Miner [7]: means enhanced classifier for data streams with novel class miner. This technique uses multiclass framework for novel class detection problem. This approach does not consider the issue of recurring classes. It is based on time constraints. So when a class is disappears for a long time and reappears again ECS Miner considered it as a novel class. It increases the error rate in the classification process.

4) Class Based Ensemble Technique: Class based ensemble technique [8] considers the data as classes and each class has a label to identify it. The class based ensemble, creates an ensemble of models for each class and each such model is called a micro-classifier. The micro-classifier uses the training data for the classification process and it is called as base learners. The proposed technique maintains this base learner and it will be updated whenever a change occurs. When a new class arrives, it is given to micro-classifier. The instances belonging to the class will be far from the existing class instances and will be closer to other novel class instances. These new instances are combined to form a new class. This technique uses cohesion and separation property. The accuracy of the micro - classifier is tested using test data.

V. CHALLENGES

Data streams show several unique properties: such as concept-drift, and concept-evolution. Concept-drift occurs in data streams when the concept of data changes over time. Concept-evolution happens when new classes evolve in streams. Each of these properties adds a challenge to data stream mining.

A. Accuracy

The accuracy of the output of many classifiers is very sensitive to concept drifts in the evolving streams. At the same time, one does not want to remove extra parts of the stream, when there is no concept drift. So methods need to be designed to decide which part of the stream need to be used for the classification process.

B. Efficiency

The process of building classifiers is a complex computational task and the update of the model due to concept drifts is a complicated process. This is especially relevant in the case of high speed data streams.

C. Robustness

Ensemble based classification has traditionally been used in order to improve robustness. The key idea is to avoid the problem of over fitting of individual classifiers. However, it is often a challenging task to use the ensemble effectively because of the high speed nature of the data streams. Many methods have been proposed to address some of these issues; they are often unable to address these issues simultaneously. Classification techniques have more importance in the research field due to the significance of their applications. A variety of methods such as cluster based method, decision trees, ECS Miner, class based ensemble method are used for the classification problem. These techniques have been designed to build classification models. Furthermore, the classification problem needs to be re-designed in the context of concept drift.

VI. CONCLUSIONS

A data stream is an ordered sequence of instances that is used in many applications. Data stream mining is the way toward retrieving useful knowledge from continuous data records. Classification is a data mining technique used to find the group membership for data instances. The classification methods for detecting novel class in data streams are explained. Classification methods are commonly strong in modeling interactions. Each classifier is different from one another. The data classification technique includes learning and classification. In learning process, the training data are analysed by classification algorithm. In classification test data are used to estimate the accuracy of the classification rules. Two key mechanisms of the novel class detection technique are outlier detection, and identifying novel class instances. They are the prime cause of high error rates in the classification. The class based ensemble technique is superior to other data stream classification techniques because of its ability to identify novel class. It also overcomes misclassification problem. However, stream classification is still faces number of challenges.

VII. ACKNOWLEDGEMENT

I would like to extend my gratitude to PG coordinator Salitha M K (Assistant Professor, Department of Computer Science and Engineering) and my guide Joshy Thomas Jose (Assistant Professor Department of Computer Science and Engineering), for their valuable guidance and suggestions. I would also thank my institution and my faculty members without whom this work would have been a distant reality.

REFERENCES

[1] T. Al-Khateeb, M. M. Masud, L. Khan, C. Aggarwal, J. Han, and B.Thuraisingham, “Stream classification with recurring and novel class detection using class-based ensemble,” in Proc. IEEE 12th Int.Conf. Data Mining, 2012, pp. 31–40.

[2] C. C. Aggarwal, J. Han, J. Wang, and P. S. Yu, “A framework for on-demand classification of evolving data streams,” IEEE Trans. Knowl. Data Eng., May 2006, vol. 18, no. 5, pp. 577–589.

[3] J. Gama, I. Zliobaite, A. Bifet, M. Pechenizkiy, and A. Bouchachia,“A survey on concept drift adaptation,” ACM Comput. Surveys, 2013, vol. 46, no. 4, pp. 44.

[4] G. Hulten, L. Spencer, and P. Domingos, “Mining time-changing data streams,” in Proc. ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining, 2001, pp. 97–106.

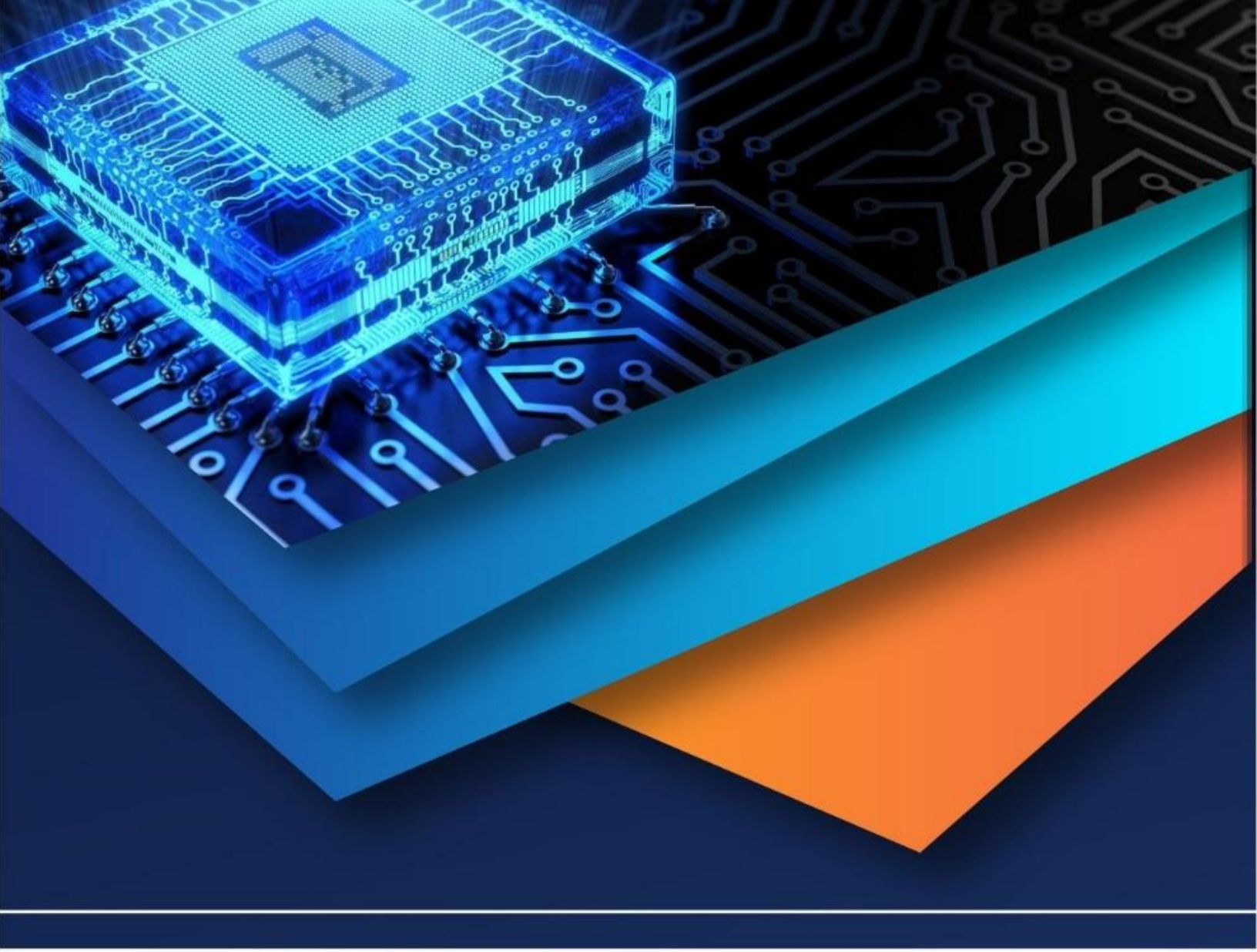
[5] E. J. Spinosa, A. P. de Leon F. de Carvalho, and J. Gama, “Cluster based novel concept detection in data streams applied to intrusion detection in computer networks,” in Proc. ACM Symp. Applied Comput., 2008, pp. 976–980.

[6] P. Li, X. Wu, and X. Hu, “Mining recurring concept drifts with limited labeled streaming data,” ACM Trans. Intell. Syst. Technol., Feb. 2012, vol. 3, no. 2, pp. 29:1–29:32.

[7] M. M. Masud, J. Gao, L. Khan, J. Han, and B. M. Thuraisingham, “Classification and novel class detection in concept-drifting data streams under time constraints,” IEEE Trans. Knowl. Data Eng., Jun.2011, vol. 23, no. 1, pp. 859–874.

[8] Tahseen Al-Khateeb, Mohammad M. Masud, Khaled Al-Naami,Sadi Evren Seker, Ahmad M. Mustafa, Latifur Khan,Zouheir Trabelsi, Charu Aggarwal, Jiawei Han, "Recurring and Novel Class Detection using Class-Based Ensemble for Evolving Data Stream", IEEE Trans. Knowl. Data Eng, October 2016.Vol. 28, NO. 10.

[9] Moharned Medhat Gaber, Arkady Zaslavsky and Shonali Krishnaswamy,"A Survey of Classification Methods In Data", 2007.



10.22214/IJRASET

45.98

IMPACT FACTOR: 7.129

IMPACT FACTOR: 7.429

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY