ISSN (Print) : 2319-5940
ISSN (Online) : 2278-1021

*International Journal of Advanced Research in Computer and Communication Engineering*
*Vol. 2, Issue 9, September 2013*

# A New Voting Method to Novel Class Detection Using Hoeffding Option Tree

Darshana Parikh [1], Priyanka Tirkha [2]

Student ofM.E., CSE, Sri Balaji College of Engg & Tech, Jaipur, Rajasthan, India [1]

Assistant Professor, CSE, Sri Balaji College of Engg. & Tech, Jaipur, Rajasthan, India [2]

**Abstract: D**ata Stream mining is a process of extracting knowledge structure from continuous and rapid data records. Data stream size is extremely large. It's a continuous flow of data So major problem related to data stream is infinite length, concept evolution and concept drift. Novel class detection is very interesting topic in a data stream mining. We can detect novel class using classification and clustering. Currently mostly uses decision tree using classification. In that Hoeffding Tree is better for data stream mining. And Hoeffding Option Tree is more better than Hoeffding Tree. In our paper we can use new voting method which is different than HOTDC. In our method we can use classification . So its supervised learning. Here classes are fixed before examined data. But when continuous data come then not all data are classified. Some data are misclassified. And this class is not in universal existing class then its known is a novel class. In our method when this type of class is detected then model is trained. So no require to collect those type of data. So when again this type of instance is come then its classified in that class. Using that method we can release from problem concept drift, concept evolution and infinite length.

**Keywords:** Novel Class, Hoeffding Option Tree, Outliers, Recurring Class, Hoeffding Bound

## I. INTRODUCTION

**Data** Mining is the practice of automatically searching large store of data to discover patterns and trends that go beyond simple analysis. Data Mining process is called "discovery," of lookng in a data warehouse to find hidden patterns without a predetermined idea about what patterns may be. So Data Mining is also known as a knowledge Discovery in Data(KDD). Data Mining is used in games, buisness,science and engineering, human rights and also in medical.In Data Mining variety of different techniques used like aritificial intelligence, neural networks, Decision Tree etc.[1]

## II. DATA STREAM MINING

Data Stream means continuous flow of data. Example of data stream include computer network traffic, phone conversation, ATM transaction, Web Searches and Sensor data. Data Stream Mining is a process of extracting knowledge structure from continuous, rapid data records.[2]Its can be considered as a subfield of data mining. Data Stream can be classified into online streams and offline streams. Online Data stream mining used in a number of real world applications, including network traffic monitoring, intrusion detection and credit card fraud detection. And offline data stream mining used in like generating report based on web log streams. Characteristics of data stream is continuous flow of data. Data size is extremely large and potentially infinite. It's not possible to store all data. But major problems related to data stream mining : Infinite length, concept evolution and concept drift. Infinite length means data stream have a infinite length so require infinite length storage and training time.[3] Concept evolution means developing novel class and concept drift means data changes over time. For our thesis main topic on concept evolution emergence of novel class. Novel class does not exist if we assume total no of classes are fixed. But some time data stream classification problem occur like intrusion detection, text classification and fault detection. So this assumption is not valid for real streaming environment. When new classes may be evolve at any time. Most existing data stream classification technique ignore this important aspect of stream data is the arrival of a Novel Class. Concept evolution solve the problem of infinite length and concept drift.

## III. NOVEL CLASS DETECTION

Novel class detection is major concept of concept evolution. In data stream classification assume that total no of classes is fixed but not be valid in a real streaming environment. When new class may evolve at any time. Most existing data stream classification technique ignore this important aspect of data stream data is arrival of a novel class.[3]

**Example.**

Classification rules:

R1. if ($x$ > $x_1$ and $y$ < $y_2$) or (x < $x_1$ and y < $y_1$) then class = +

R2. if ($x$ > $x_1$ and $y$ > y2) or (x < $x_1$ and y > $y_1$) then class = -

Existing classification models misclassify **novel class** instances

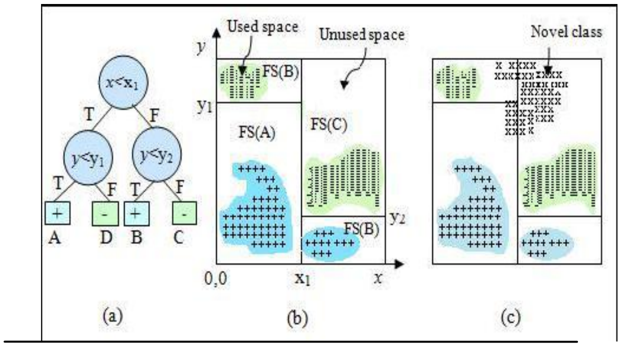

**Fig 1: (a) Decision Tree (b) Corresponding feature space partitioning where FS(X) denote the feature space defined by a leaf node X the shaded area shows the used space in each partition. (c) Novel class (x) arrives in unused space. [6]**

## IV. RELATED WORK

**Mine class** which stands for mining novel class in data streams with base learner K-NN and decision tree. In previously assume that before finding novel class one class is normal others are novel class. Mining class provide multi-class classification framework. Data stream classifier are divided into two category single model and ensemble model. Single model incrementally update a single classifier and effectively respond to concept drifting so that reflects main concept in data stream. Ensemble model use combination of classifiers with a aim of creating and improved composite model.[1]

Actminer applies on ensemble classification technique but used for limited data problem and addressing other three problems so reducing cost. This technique extends from Mine Class. But in this technique dynamic feature set problem and multi label classification in data stream classification.[3]

ECSMiner technique provides multiclass framework for novel class detection problem that can distinguishes between different classes of data and emergence of a novel class. ECSMiner can not distinguish between novel class and recurring class.[2]

SCANR technique used by both primary ensemble and auxiliary ensemble method. So Error rate is reduced than ECSMiner. And also it can distinguish between novel class and recurring class.[1]

Decision Tree approach is used for when data stream is continuously changes. In that calculate threshold value based on number of data points between each leaf node in a tree and training data set and cluster the data points of training data set based on similarity of attributes.[4]

## V. LEARNING ALGORITHM

HOEFFDING TREE

**Hoeffding tree** algorithm is a state-of-the-art method including decision trees from data streams. In a tree mainly we can tree will be evaluate form left order. But in Hoeffding tree we can use hoeffding bound for create a tree. Hoeffding tree store data stream only once. After that update tree. The Hoeffding bound states that with probability 1 – δ the true mean of a random variable of range R will not differ from estimated mean after n independent observation by more than:

$$\epsilon = \sqrt{R^2\ln(1/\delta) / 2n}$$

This bound is useful because it holds true regardless of the distribution generating values.[9]

## VI. OPTION TREE

Option tree possible for an example to travel down multiple paths and arrive at multiple leaves. This is possible by option nodes of the tree. Option node splits different paths with different ways. [7]Making a decision with option tree involves combining the prediction of applicable leaves into final result. Option tree over traditional ensemble method is that more flexible to representation and save the space. Suppose in normal tree if 100 mostly identical large tree and single leaf than require space for each leaf. But in option tree require hundred time less space required. Option tree contain many option at many levels can be complex, but humans may not be confused as limited no of option nodes in small and simple option tree. Its increase accuracy.[10]

## VII. HOEFFDING OPTION TREE

The Hoeffding tree based learner are less accurate when several attribute appear equally discriminative. At that time we can not take any decision based on Hoeffding bound. For that we can use user defined threshold which implicitly specifies amount of the data that has to be observed in order to decide one of the competitive attribute. [6] So reduced accuracy because of delay. We can solve this problem using option tree including option node with splitting nodes. Option tree does not increase the complexity of the tree. Option tree mainly used when ambiguity occurs. In this figure this tree is regular decision tree including option nodes in form of rectangles. At this node multiple tests will be applied and travel down multiple paths and multiple leaves.[9]

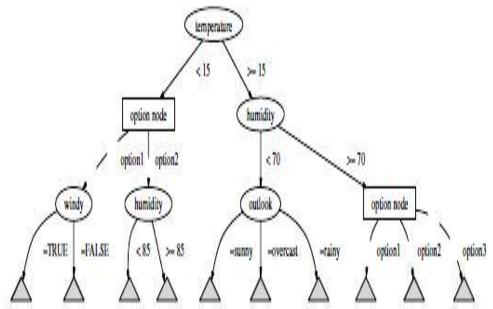

**Fig-2 An Option Tree [6]**

## VIII. Limitation Of Existing System

Above We can see the methods like Mining algorithm, ECSMiner, SCANR this methods are based on Clustreing. Based on clustering methods for collecting potential novel instances so memory is required to store that so memory overhead occurs. Another disadvantage is that using clustering method first find centroid for that CPU overhead occurs. And also incremental so time overhead occurs. And also not possible classify streamed data continuously. Because streamed data continuously come and classification become continuous task. [1][2][3][4]

## IX. Proposed System

So for novel class detection new method is decision tree which is based on classification. Hoeffding tree is most powerful for streamed data. And Hoeffding Option Tree for classification is better than Hoeffding tree. Because option are naturally dealt with equally discriminative attribute. Also in our proposed system model is trained as when potential novel instance is found . So not requiring to collect novel instances. So memory overhead not occur. Also classification process start from latest novel instance not require to find centroid . So improve CPU timing. Here we can define some terms which are useful for proposed system.

**Current Existing class**: Let M be a current primary ensemble. A class c is current existing class. If any model $M_i$ ε M has been trained with class c.[1]

**Universal existing class :** Let $M_g$ ε { M1,M2 …. $M_i$}. A class c is a universal existing class if at least one of the models $M_i$ ε $M_g$ has been trained with class c.[1]

I. **Novel Class :** A class is c is a novel class if it is not universal existing class. In oter words class c is called novel class if it is never appeared before in the stream.[1]

**Outlier** : A data object that deviates significantly from the normal objects as if it were generated by different mechanism.[1]

In previous decision tree method first we decide how many classes are available. And then when data is come then tree is generated. But its occur when some data is not classified. Its misclassified. After generation of the tree check misclassified data is more than classified data. If its true then again create a new class model is trained then new tree will be generated. In our proposed system we use Hoeffding option tree which is better than Hoeffding tree. Option are naturally way to deal with equally discriminative attribute. And also Hoeffding option tree is better method for classification for stream data.

In our approach first decide the classes. When data is come than according to that tree will be created. But in that some data are misclassified. But no need to collect misclassified instances. When we detect single potential instance then model is trained , classify that instance and create a new class. So when those type of instances come that then classify in that class. So no require to collecting misclassified instances. So memory is not wasted. And clustering not require so CPU timing is not wasted.

## X. Proposed Algorithm

```
For each instance I in the stream
    Classify(hot , I)
    If ( I is misclassified)
        If( class of I is already appeared in model) then
            I is misclassified instance
        Else if (class of I is not in model) then
            I is novel instance
        End if
    End if
    If (I is novel instance)
        Train the model with latest instance I
        Classify(hot,i)
    End if
End For
End
```

## XI. EXPERIMENTAL ANALYSIS

### DATA SET DESCRIPTIONS

| Data Set | No Of Attributes | No. Of Instances | No. Of Class | Type |
|---|---|---|---|---|
| Powersupply | 3 | 29928 | 24 | numeric |
| Cedit-g | 21 | 1000 | 02 | Nominal |
| Vowel | 14 | 990 | 11 | Nominal |
| Votes | 17 | 435 | 02 | Nominal |
| Vehicle | 19 | 846 | 04 | Numeric |

## XII. RESULTS:

### COMPARISON HOT AND HOTND

| Data set | HOT | | HOTND | |
|---|---|---|---|---|
| | Accuracy | Time | Accuracy | Time |
| Powersupply | 11.00 | 0.27 | 12.20 | 0.47s |
| Cedit-g | 77.60 | 0.03s | 84.90 | 0.06s |
| Vowel | 74.95 | 0.08s | 90.30 | 0.09s |
| Votes | 92.87 | 0.02s | 93.79 | 0.02s |
| Vehicle | 47.28 | 0.05s | 60.99 | 0.06s |

Here this algorithm is implemented in Java which is most popular open source platform independent language. Here we use MOA tool which have capability to experiment on stream data classification with HOT and others. Here Hot has been adapted from MOA. And here evaluate method is evaluate prequential which is used in HOT. In comparision table we can see accuracy of HOTND is higher than HOT. Because Hot can not find novel instances. Its misclassify those data. HOTND detect novel instance and also classify. So accuracy of HOTND is higher than HOT. But here we can see time is higher taken by HOTND than HOT. Because HOTND detect novel instance and than classify it so time is more require than HOT.

## XIII. ADVANTAGES

HOTND reduced memory requirement because no requirement for collecting novel instances. Also improve accuracy because start to train classifier when a first novel instance found. Also CPU timing reduced because does not require clustering. So not require to find centroid. This algorithm solve the problem of concept evolution, infinite length and concept drift.

## XIV. CONCLUSION

In this paper we introduce new voting method to detect novel class using hoeffding option tree in concept drifting data stream classification which builds a decision tree from data stream. Here we can train a model when potential novel instance is found. Not require to collect misclassified instances. So do not require to further classification. Timing and accuracy is improved. Here we can compare with HOT and HOTND(New voting method). But Hot did not classify novel instances so time require is smaller than HOTND. But accuracy of HOTND is higher than HOT.

## REFERANCES

[1] Mohammad M Masud, Tahseen M, Al-khateeb, Latifur Khan, Charu Aggrawal, Jing Gao, Jiawei Han and Bhawani Thuraisinghum Detecting Recurring and Novel classes in Concept Drift Data Streams icdm, pp. 1176-1181, 2011 IEEE 11th International Conference On Data Mining.

[2] S.Thanngamani DYNAMIC FEATURE SET BASED CLASSIFICATION SCHEME UNDER DATA STREAMS *International Journal Of Communication And Engineering Volume 04 – No .04, Issue:01 March-201.*

[3] Mohammad M. Masud, Jing Gao, Latifur Khan, Jiawei Han, Bhavani Thuraisingham Classification And Novel Class Detection In Data Stream With Active Mining M.J.Zaki etal.(Eds.): PAKDD 2010, Part II,LNAI 6119, pp.311-324 Springer- Verlag Berlin Heidelberg 2010.

[4] Amit Biswas, Dewan Md. Farid and Chowdhary Mofizur Rahman A New Decision Tree Learning Approch For Novel Class Detection In Cocept Drifting Data Stream Classification JOURNAL OF COMPUTER SCIENCE AND ENGINEERING, VOLUME 14, ISSUE 1, JULY 2012.

[5] S.PRASANNALAKSHMI,S.SASIREKHA INTERGATING NOVEL CLASS DETECTION WITH CONCEPT DRIFTING DATA STREAMS *International Journal Of Communication And Engineering Volume 03, No. 03, Issue:04 March 2012.*

[6] JIGNASA N. PATEL, SHEETAL MEHTA Detection Of Novel Class With Incremental Learning For Data Streams International Journal Of Research in Modern Engineering and Emerging Technology Vol.1, Issue:3 April-2013.

[7] Geoffrey Holmes, Richard Kirkby, and Bernhard P Fahringer Mining Data Stream Using Option Trees(revised edition 2004).

[8] Nabil M. Hewahi, Motaz K. Saad Class Outlier Mining: Distance Based Approch International Journal Of Intelligent Technology, Vol. 2, No. 1, pp 55-68, 2007

[9] Pedro Domingos, Geoff Hulten Mining High-Speed Data Streams in proceeding of the 6th ACMSIGKDD International Conference On Knowledge Discovery and Data Mining, pp.71-80, ACM, August-2000

[10] Ron kohavi, Clayton Kunj Option decision tree with majority votes in international conference of machine learning pages 161-169, 1997.