IN DATA STREAMS USING CLASSIFICATION AND CLUSTERING DIFFERENT TECHNIQUES TO FIND NOVEL CLASS

Darshana Parikh[1], Priyanka Tirkha[2]

[1, 2]*Department of Computer Science & Engineering, Sri Balaji College of Engineering & Technology*
darshana_shruti@yahoo.co.in , tirkhapriyanka@gmail.com

Abstract

Data stream mining is a process of extracting knowledge from continuous data. Data Stream classification is major challenges than classifying static data because of several unique properties of data streams. Data stream is ordered sequence of instances that arrive at a rate does not store permanently in memory. The problem making more challenging when concept drift occurs when data changes over time Major problems of data stream mining is : infinite length, concept drift, concept evolution. Novel class detection in data stream classification is a interesting research topic for concept drift problem here we compare different techniques for same.

Index Terms— *Ensemble Method, Decision Tree, Novel Class, Option Tree, Recurring class*

----------***----------

1. INTRODUCTION

Data Mining is a process of extracting hidden useful information from large volume of database. Data stream is order sequence of instance that arrives any time does not permit to store permanently in memory. 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 looking 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, business, science and engineering, human rights and also in medical. In Data Mining variety of different techniques used likes artificial intelligence, neural networks, Decision Tree etc.

Data mining process has two major functions: classification and clustering.[1],[3],[5] In a data stream classification assumed that total no of classes are fixed. Its not valid for real environment when new classes may involve. The goal of data mining classifiers is predict the class value or unseen instances 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. [5]

2. DATA STREAM MINING

Data Stream means continuous flow of data. Example of data stream includes computer network traffic, phone conversation, ATM transaction, and Web Searches and Sensor data. Data Stream Mining is a process of extracting knowledge structure from continuous, rapid data records. 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 are continuous flow of data. Data size is extremely large and potentially infinite. It's not possible to store all data Data stream classification three major problems occurred.

- Infinite Training Data
  - Can't store or use all historical data for training.
- Concept drift
  - Data changes over time. Historical training data built a model on those data which are outdated.
- Novel class
  - Novel class may appear over time. Old classes become obsolete (out dated).[5]

Data stream have infinite length multi pass learning algorithm can not applicable as they would required infinite storage. Concept drift occurs when data changes over time. Another major problem is ignored by state of art data stream classification techniques which is concept evolution that means emergence of novel class. Assume that total no of classes is fixed. But in real data stream classification problems such as intrusion detection, text classification and fault detection Novel class may appear at any time in a stream. So all novel class instance go undetected until novel class manually detected by experts.

When a new class emerges than classifier misclassify those instances because classifier is not trained with those class . Data stream classifiers are divided into two models. 1) Single model. 2) Ensemble model. In single model incrementally update a single classifier effectively updates concept drift. Ensemble model use a combination of classifiers with the aim of improve composite model. A fixed sized ensemble is used to classify data streams and detect novel class. [1]

In this primary ensemble M and auxiliary ensemble to approaches used. In that first stream is divided into equivalent chunks. Data points in latest chunk first classified using ensemble. But when data points between chunks become labeled that chunk is used for training a classification model. Number of methods in each ensemble is fixed , newly trained model replaces existing model in each ensemble. Each incoming unlabeled instance is first classified by outlier detection module of primary ensemble to check its outlier or not. If it is not an outlier than it is classified as an existing class using majority voting in classifiers in primary ensemble. If it is an outlier then it's called primary outlier otherwise check by auxiliary ensemble. It is called secondary outlier and temporary stored in a buffer. And novel class techniques invoked. If novel class found than tagged with novel class instance. Here so many techniques

For novel class detection. [1] Explain in section Iv.

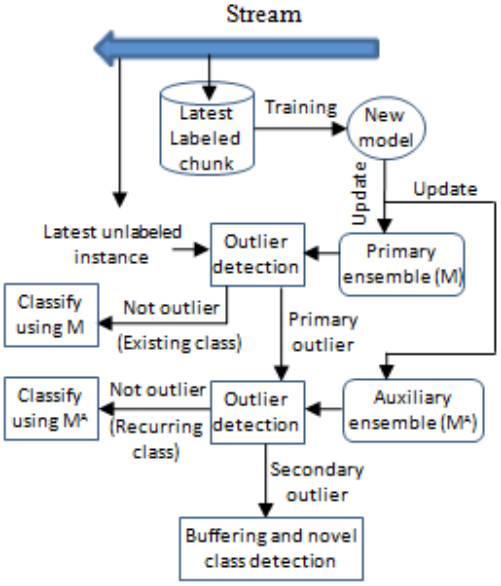

Fig1. Architecture of SCANR technique

3. 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 > x1 and y < y2) or (x < x1 and y < y1) then class = +

R2. If (x > x1 and y > y2) or (x < x1 and y > y1) then class = -

Existing classification models misclassify novel class instances

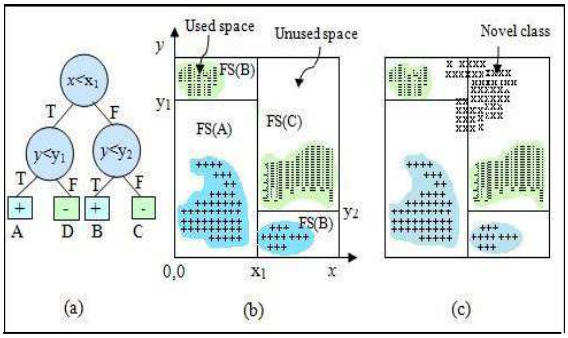

Fig 2: (a) Decision Tree (b) Corresponding feature space partitioning where FS(X) denote the feature space defined by a leaf node X the shaded area shows the used space in each partition. (c) Novel class (x) arrives in unused space.

4. DIFFERENT TECHNIQUES FOR DETECT NOVEL CLASS

4.1 Actminer

Actminer applies an ensemble classification technique but used for limited labeled data problem and addressing the other three problem so reducing the cost. Actminer is extends from mine class. Actminer integrates with four major problem concept drift, concept evolution, novel class detection, limited labeled data instances. But in this technique dynamic feature set problem and multi label classification in data stream classification. [3]

4.2 ECSMiner

ECSMiner means enhanced classifier for data streams with novel class miner. This Technique provides "multiclass" framework for novel class detection problem that can distinguishes between different classes of data and emergence of a novel class. This technique considers time constraints. These techniques applied on two different classifiers: decision tree, K-NN nearest neighbor. When decision tree is used as a classifier, each training data chunk is build a decision tree. When K-NN is used, each chunk is used for classification model. [2]

In [1] "Recurring class" is a special case of concept evolution. It occur when class reappears after long disappearance of the stream. ECSMiner identifies recurring class as a novel class.

4.3 SCANR

SCANR which stands for "Stream Classifier And Novel class And Recurring Class detector". In that each incoming instance first checked by primary ensemble M to see if its an outlier for M(P-outliers).P-outliers further passed to auxiliary ensemble for further check. If it is not P-outliers then its normally classified otherwise stored in a buffer for further analysis. Finally buffer is checked for novel class .Novel class check is done sparingly to reduce cost and redundancy. If we compare ECSMiner and SCANR then ERR rate of ECSMiner is more than SCANR because ECSMiner can not distinguish between novel class and recurring class. FNew rate is more than SCANR because it take small time for classification.[1]

4.4 Decision Tree

A new decision tree learning approach for novel class detection. In this builds a decision tree from data stream which continuously update. Calculate threshold value based on ratio of percentage of data points between each leaf node in a tree and the training dataset and cluster the data points of training data set based on similarity of attributes. If number of data points classify at a leaf node increases than the threshold value increases then novel class arrived. ID3 technique builds a decision tree using information theory. ID3 choose a splitting attribute from a dataset with the highest information gain. C4.5 is a successor of ID3 through gain ration. For splitting purpose, C4.5 use largest gain ration that ensures larger than average information gain. CART (Classification and Regression Tree) is a process of generating binary tree for decision making. CART handles missing data and pruning strategy. SPRINT (Scalable Parallelizable Induction of Decision Tree) algorithm uses impurity of function called gini index to find best split. In this they introduce decision tree classifier based novel class detection in concept drift data stream classification which builds a decision tree from data.

4.5 Hoeffding Option Tree

Hoeffding trees are state-of-the-art for processing high speed data streams. Hoeffding option tree is a regular Hoeffding tree containing additional option nodes that allows several tests to be applied, leading to multiple hoeffding trees as multiple paths. When training model on a data stream it is important to make a single scan of data as quickly as possible.[6] Option tree represents middle ground between single trees and ensembles. They are capable of producing useful and interpretable, additional model structure without consuming too many resources. Option tree consists of a single structure that efficiently represents multiple trees. It can travel down on multiple paths of the tree and different options. [7]

CONCLUSIONS

Most challenging task in data stream to detect a novel class. Here we have studied about different techniques for detect novel class using classification and clustering. But in classification decision tree is very easy approach to find novel class. So we can use different algorithm for decision tree and finding novel class. Also we can change voting technique and move towards.

REFERENCES

[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).