Detection of Novel Class with Incremental Learning for Data Streams

JIGNASA N. PATEL
Student of M.E (C.E.)
Parul Institute of Engineering & Technology,
Waghodia, Vadodara, Gujarat, India

SHEETAL MEHTA
Assistant Professor,
CSE Department
Parul institute of Engineering & Technology,
Waghodia, Vadodara, Gujarat, India

Abstract:
Data stream mining is the process of extracting knowledge from continuous arrival of data. Data stream can be viewed as a sequence of relational touples arrives continuously at any time. Classification of data stream is more challenging task due to three major problems in data stream mining: Infinite length, Concept-drift, Arrival of novel class. Novel class detection in stream data classification is interesting research topic and researches available for concept drift problem but not attention on the Novel class detection. In this paper, we propose a new approach for detecting novel class in concept drifting DataStream classification using decision tree classifier that can determine whether data instance belongs to existing class or a novel class. The proposed approach is to find Novel Class instances using Hoeffding Option tree (HOTDC) and from training data points, which continuously updates with recent instances so that the tree represents the most recent concept in data stream. The experimental analysis on datasets from UCI machine learning repository proved that the proposed approach can detect novel class in concept drifting data stream classification problems.

Keywords: *Data stream, Incremental learning, Hoeffding Option Tree, Novel Class*

1. Introduction

Data mining is the process of extracting hidden useful information from large volume of database. A data stream is an ordered sequence of instances that arrive at any time does not permit to permanently store them in memory. Data mining process has two major functions: classification and clustering. Data stream classification is the process of extracting knowledge and information from continuous data instances. The goal of data mining classifiers is to predict the class value of a new or unseen instance, whose attribute values are known but the class value is unknown [1].Classification maps data into predefined that is referred to a supervised learning because the classes are determined before examining the data and that analyses a given training set and develops a model for each class according to the features present in the data. In clustering class or groups are not predefined, but rather defined by the data alone. It is referred to as unsupervised learning.

There are three major problems related to stream data classification [2]. It is impractical to store and use all the historical data for training

There may be concept-drift in the data, meaning, the underlying concept of the data may change over time.

In data stream classification most of the existing work related to infinite length and concept drift here we focus on the novel class detection. Stream classification problems, such as intrusion detection, text classification, fault detection, novel classes may arrive at any time in the continuous stream. There are many approaches to develop the classification model including decision trees, neural networks, nearest neighbor methods and rough set-based methods [4].The data stream classifiers are divided into two categories: single model and ensemble model [1]. Single model incrementally update a single classifier and effectively respond to concept drifting so that reflects most recent concept in data stream. Ensemble model use a combination of classifiers with the aim of creating an improved composite model, and also handle concept drifting efficiently. The traditional tree induction algorithm is that they do not consider the time in which the data arrived. The incremental classifier that reflects the changing data trends effective and efficient so it is more attractive. Incremental learning is an approach to deal with the classification task when datasets are too large or when new examples can arrive at any time [5]. Incremental learning most important in applications where data arrives over long periods of time and storage capacities are very limited. In [7] author Defines incremental tasks and incremental algorithms as follows:

Definition 1: A learning task is incremental if the training examples used to solve it become available over time, usually one at a time.

As per [8] the learning to be one that is: Capable to learn and update with every new data (labeled or unlabeled), Will use and exploit the knowledge in further learning, Will not rely on the previously learned knowledge, Will generate a new class as required and take decisions to merge or divide them as well Will enable the classifier itself to evolve and be dynamic in nature with the changing environment. Decision tree that provide the solution for handling novel class detection problem. ID3 is very useful learning algorithm for decision tree.C5.0 algorithm improves the performance of tree using boosting. Hoeffding tree containing additional option nodes. Option tree represent middle ground between Incremental and Ensemble approach. HOT that control tree growth and determine number of option to explore [13].

2. Novel Class Detection

Novel class detection in stream data classification is interesting research topic and researches available for concept drift problem but not attention on the Novel class detection. Data stream classification and novelty detection recently received increasing attention in many practical real-world applications, such as spam, climate change or intrusion detection, where data distributions inherently change over time[6]. Ensemble techniques maintain a combination of models, and use ensemble voting to classify unlabeled instances. As per [6] In 2011, Masud et al. proposed a novelty detection and data stream classification technique, which integrates a novel class detection mechanism into traditional mining classifiers that enabling automatic detection of novel classes before the true labels of the novel class instances arrive. In [9], [10] author gives the definition of the existing class and Novel class.

Definition 1 (*Existing class and Novel class*): Let L is the current ensemble of classification models. A class c is an existing class if at least one of the models Li ∈ L has been trained with the instances of class c. Otherwise, c is a novel class in 12 points boldface italic and capitalize

the first letter of the first word only. Do not underline any of the headings, or add dashes, colons, etc.

In [10] show the basic idea of novel class detection using decision tree in Figure 1. That introduces the notion of used space to denote a feature space occupied by any instance, and unused space to denote a feature space unused by an instance.

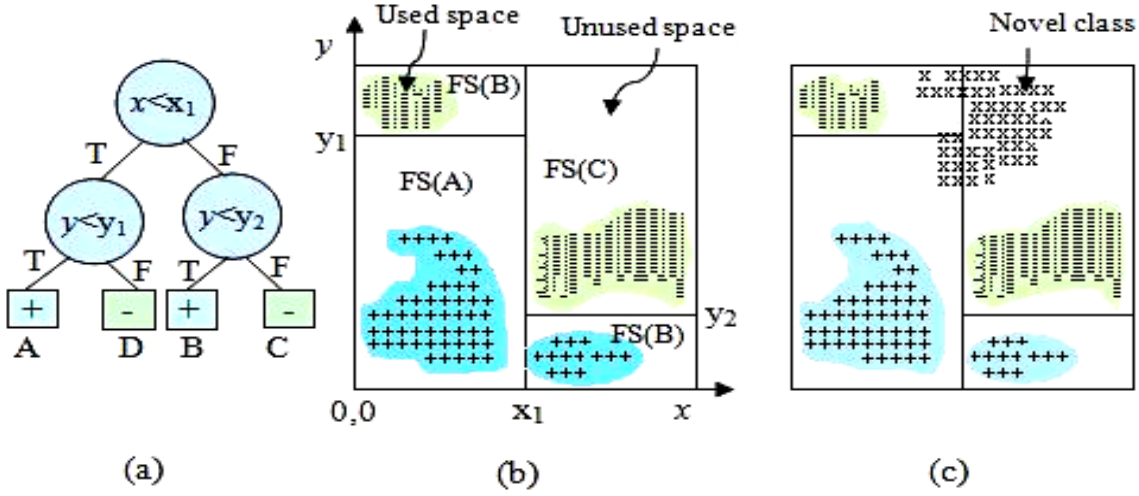

Fig. 1 (a) A decision tree, (b) corresponding feature space partitioning where FS(X) denotes the Feature space defined by a leaf node X The shaded areas show the used spaces of each partition.(c) A Novel class (denoted by x) arrives in the unused space.

First, the classifier is trained such that an Inventory of the used spaces is created and saved. This is done by clustering and saving the cluster summary as "pseudo point" (to be explained shortly). Secondly, these Pseudo points are used to detect outliers in the test data, and declare a novel class if there is strong Cohesion among the outliers.

3. Related Work

In [10] author describe "Mine Class", which stands for Mining novel Classes in data streams with base learner K-NN (K-nearest neighbor) and decision tree. Novelty detection is also closely related to outlier/anomaly detection techniques. The main difference with this outlier detection is that here primary objective is novel class detection, not outlier detection. Outliers are the by-product of intermediate computational steps in Novel class detection algorithm. Recent work in data stream mining domain describes a clustering approach that can detect both concept-drift and novel class and assumes that there is only one 'normal' class and all other classes are novel. Thus, it may not work well if more than one class is to be considered as 'normal' or 'non-novel'. Mine Class addresses the concept evolution problem on a multi-class classification framework. Mine Class does not address the limited labeled data problem, and requires that all instances in the stream be labeled and available for training.

In [9] Act Miner applies an ensemble classification technique by addressing the limited labeled data problem. Act Miner extends Mine Class, and addresses the Limited labeled data problem in addition to addressing the other three Problems thereby reducing the labeling cost, but it is not applicable to multi-class classification. In [11] author describes ECS Miner for Novel class detection. Novel class detection using ECS Miner offers a "multiclass" framework for the

novelty detection problem that can distinguish between different classes of data and discover the emergence of a novel class. This technique is a nonparametric approach, and therefore, it is not restricted to any specific data distribution. "ECS Miner" (pronounced like Ex Miner).This technique on two different classifiers: decision tree and k-nearest neighbor. When decision tree is used as a classifier, each training data chunk is used to build a decision tree. K-NN strategy would lead to an inefficient classification model, both in terms of memory and running time.

In [12] author proposed a *recurring class* is a special case of concept-evolution. A *recurring class* is a special and more common case of concept-evolution in data streams. It occurs when a class reappears after long disappearance from the stream. ECS Miner identifies recurring classes as novel class. Each incoming instance of data stream is first check by primary ensemble if it is outlier called it primary outlier (P-outlier) than again check through auxiliary ensemble if it is outlier than called *secondary outlier(S-outlier)*, and it is temporarily stored in a buffer for further analysis. When there are enough instances in the buffer, the *novel class detection* module is invoked.

4. Learning Algorithm

Data mining is the process of finding hidden information and patterns in a huge database. Data mining algorithms have two major functions: classification and clustering. Classification maps data into predefined groups or classes. It is often referred to a supervised learning because the classes are determined before examining the data. Classification creates a function from training data. On the other side, clustering is similar to classification except that the groups are not predefined, but rather defined by the data alone. It is alternatively referred to as unsupervised learning.

4.1 Hoeffding Option Tree

As per describe in [13] Hoeffding trees are state-of-the-art for processing high-speed data streams. Hoeffding Option Trees is a regular Hoeffding tree containing additional *option* nodes that allow several tests to be applied, leading to multiple Hoeffding trees as separate paths. When training a model on a data stream it is important to make a single scan of the data as quickly as possible. Hoeffding trees achieve this by accumulating sufficient statistics from examples in a node to the point where they can be used to make a sensible split decision. The sufficient statistics are beneficial for both tree growth and prediction as they can be used to build Naive Bayes models at the leaves of the tree that are more accurate than majority class estimates. Option trees represent a middle ground between single trees and ensembles. They are capable of producing useful and interpretable, additional model structure without consuming too many resources. Option trees consist of a single structure that efficiently represents multiple trees. A particular example can travel down multiple paths of the tree, to different options.

4.2 Option Node

Figure 2 is an example of what the top few levels of an option tree can look like. The tree is a regular decision tree in form except for the presence of option nodes, depicted in the figure as rectangles. At these nodes multiple tests will be applied, implying that an example can travel down multiple paths of the decision tree, and arrive at multiple leaves. Option Tree is control tree growth, and determined a reasonable number of options to explore it. As per [14] option nodes are a natural and effective solution to the problem of dealing with multiple equally discriminative attributes (the tie problem). The additional structure of the option trees provides interesting and useful information on the ambiguity of the splits and thus on the existence of several equally relevant attributes.

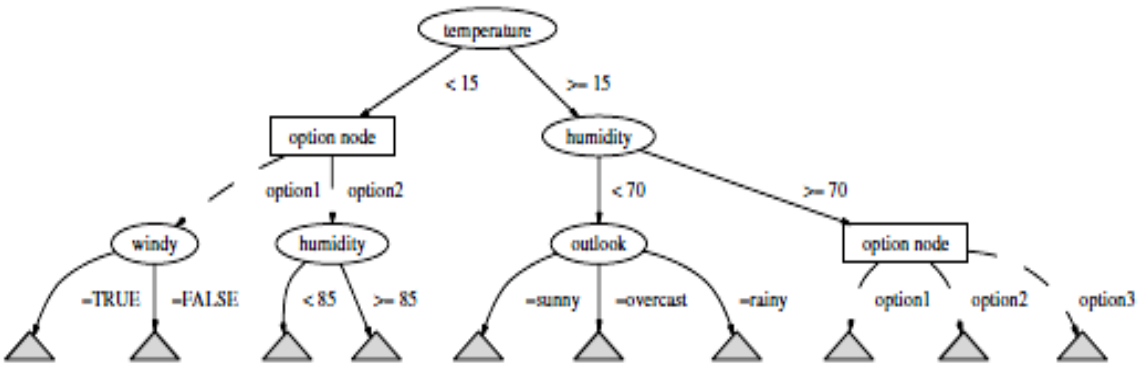

Fig. 2 An option tree

5. Proposed Algorithm (HOTDC)

1. Buf empty // Temporary Buffer
2. U empty // Unlabel data buffer
3. L empty // Label Data buffer (Training Data)
4. While (MaxInstance Variable) do
5. Xj the latest Instance in the stream
6. Classify (HOT, Xj, buf)
7. If Class label cannot predict immediately sore into U
8. Else Stored into L
9. End if
10. End While
11. While (Max Instance in U)
12. Classify (HOT, U)
13. Remaining Instances in U are Novel Instance
14. End while
15. End

6. Experimental Analysis

6.1 Dataset

Data stream mining is the process of analyzing online data to discover patterns, which uses sophisticated mathematical algorithms to segment the continuous data and evaluate the probability of future events. A set of data items called the dataset, which is the very basic concept of data mining and machine learning research.

Table 1 Data set descriptions

Dataset	No. of Attribute	No. of Instances	No. of Class	Type
Contact Lense	04	24	03	Real
Zoo	18	101	07	Real
Votes	16	435	02	Real

Dataset	No. of Attribute	No. of Instances	No. of Class	Type
Elec Norm	08	45,312	02	Real
Iris	04	150	03	Real
Waveform-5000	21	5000	03	Synthetic
Soybean	35	683	19	Real

7. Results

We implement our algorithm in Java. The code for Hoeffding Option tree has been adapted from the Massive Online Analysis (MOA) open source repository. The tools for data pre-processing, classification, regression, clustering, association rules, and visualization. The experiments were run on an Intel Core 2 Duo Processor with 1.99 GB of RAM. There are various approaches to determine the performance of data stream classifiers. Here, the performance can most simply be measured by counting the proportion of correctly classified instances in an unseen test dataset. Table 2 shows the classification accuracy, Kappa Statics and number of novel instances that are not classified.

Table 2 Comparison of Hoeffding Option Tee and Hoeffding Option Tree with detection of Novel Instances

Dataset	Hoeffding Option Tree (HOT)			Proposed Algorithm (HOTDC)		
	Accuracy	Kappa	Novel Instance	Accuracy	Kappa	Novel Instances
Contact Lenses	70.83	50.59	-	95.48	64.48	07
Zoo	87.06	83.04	-	95.37	89.92	13
Votes	88.35	75.44	-	91.48	81.76	42
Elec Norm	85.64	71.14	-	85.91	71.69	124
Iris	95.50	92.50	-	97.60	94.13	09
Waveform 5000	79.24	68.84	-	79.72	69.40	212
Soybean	90.87	89.73	-	91.10	89.99	71

8. Conclusion

In this paper, we introduce Hoeffding Option tree classifier based novel class detection in concept drifting data stream classification, which builds a decision tree from data stream. The HOTDC (Hoeffding Option Tree with detection of Novel Class) continuously updates with new data points so that the most recent tree represents the most recent concept in data stream. The main propose of this paper is to improve the Accuracy Performance of HOT in concept drifting data stream classification problem. The HOT classifier is very popular supervised learning algorithm that has several advantages requires little prior knowledge. We tested the performance

of HOTDC on several benchmark datasets that efficiently detect novel class and improve the classification accuracy.

References

1. Comparative Computational Analysis of Drag-Reducing devices For Tractor-Trailers.
2. Cooper, Kevin R. (2005). Model and Full-Scale Wind Tunnel Tests of Second-Generation Aerodynamic Fuel Saving Devices for Tractor-Trailers Jason Leuschen Aerodynamics Laboratory, NRC, Ottawa, Canada.
3. Peterbilt Motors Company presents a white paper on Truck aerodynamics and thermal efficiency.
4. Robert M. (2006). Clarke truck Manufacturers association doe Heavy Vehicle Systems optimization Merit Review, Truck Manufacturers Program to Reduce Aerodynamic Drag.
5. School of Computing and Engineering Researchers (2008). Conference, University of Hudders field