Handling of Recurrence Concept Drift in Data Stream using Timestamp of Auxiliary Learning Model

D.SUJATHA
Dept of CSE
Sri Sai Aditya Institute of Science and Technology
Surempalem, India

M.ANIL KUMAR
Senior Asst. Professor Dept of CSE
Sri Sai Aditya Institute of Science and Technology
Surempalem, India

Abstract: - Data stream is a collection or sequence of data instances of infinite length. Stream classification or online classification is more challenging task due to speed, diversity of concept or nature, type of distribution (linear or skewed), heterogeneous of data sources, lack of re-reading of instances and possibility of recurrence. This paper focuses on the concept drift under recurrence. The major challenge in data stream is handling of high volume of data of infinite length. Classification of instances under concept drift and recurrence is more difficult due to maintenance of past classifier results. To handle this situation in more efficient manner through swapping technique followed operating system’s demand paging concept with little modifications

Keywords: Data Stream Mining, Concept Drift, Recurrence Concept Drift

I. INTRODUCTION

Concept Drift detection is one of the ever challenging problem in data stream management. It has been receiving great attention in recent researches of stream mining. Wang et al. [2011] describe that, in machine learning, “the term concept refers to the quantity that a learning model is trying to predict, i.e., the variable. Concept drift is the situation in which the statistical properties of the target concept change over time.” Applications which contain concept drift are Intrusion Detection System, Spam Filtering and credit card fraud detection.

There are different types of concept drift. First category is based on speed of change. Sensor streams are one of example of such concept drift. If sudden change of transition occurred from old to new concept, then it is treated as abrupt concept drift. If the same transition change occurrence is slow or smooth then it is treated as gradual concept drift. Concept drifts are also classified based on reason of change. They are of 2 types real and virtual concept drifts. If the valid labels of training examples are occurred at one time and has different valid labels at other time then this change is called Real Concept Drift. One more situation in data stream is recurrence of Context. This will be happened if the same concept which was already classified and out dated is repeated again in the current stream. This is due to cyclic transactions such as seasonal business transactions, market tendency, weather conditions etc. Most of the algorithms try to remember current or recent set of drifts and old drift classes are invalidated to reduce the memory occupancy and search time complexity.

II. BACKGROUND

Hulten et al. [2001] proposed concept-adapting very fast decision tree and it is a extension of very fast decision tree (VFDT) [Domingos and Hulten, 2000]. Very Fast Decision Tree for Continues attributes has been introduced by Gama et al. [2003]. Gama et al.[2004a] introduced Drift Detection Method (DDM). This method is classifier independent and detects drift in the form of context. Other method which most similar to DDM is Early DDM (EDDM) developed by Baena-García et al. [2006]. DDM detects the drift based on the controlling of classifiers error rate from time to time between two boundary values where as EDDM uses distance between errors.

Some other methods used to handle concept drift is the use of ensembles. One of such method Accuracy Weighted Ensemble (AWE) classifier has been introduced by Wang et al. [2003]. Other method is Accuracy Updated Ensemble (AUE), developed by Brzeziński and Stefanowski [2011]. Both of these methods use weight but AWE use batch of classifiers and AUE uses incremental classifiers.

Some other existing approaches deal with recurrence concept drift. FLORA3 [Widmer and Kubat, 1996] and SPLICE-2 [Harries et al., 1998] store information about the concepts and, if necessary, reuse them. Flora represents a concept description in the form of three description sets: “the set ADES (Accepted Descriptors) contains description items matching only positive examples. The set NDES (Negative Descriptors) summarizes the negative examples. PDES (Potential Descriptors) contains description items that are too general, matching positive examples, but also some negative ones. SPLICE-2 also deals with categorical attributes but the classifier training is made in batch mode.

Figure 1 Illustration of the four structural types of the drift (Ref: Indre Zliobaite "Learning under Concept Drift: an Overview", Oct 2010)

III. PROPOSED APPROACH

Data stream is a collection or sequence of data instances of infinite length. Stream classification or online classification is more challenging task due to speed, diversity of concept or nature, type of distribution (linear...
This paper focuses on the concept drift under recurrence. The major challenge in data stream is handling of high volume of data of infinite length. Classification of instances under concept drift and recurrence is more difficult due to maintenance of past classifier results. To handle this situation in more efficient manner through swapping technique followed operating system’s demand paging concept with little modifications. The figure-2 describes basic architecture of proposed approach.

Algorithm-1: Online Classification
1. Online reading of DataStream
2. Assignment of Time stamp to each instance
3. Building of some initial window of instances for some time period
4. Train the ensemble of classifiers with initial window of instances
5. Consider next time window of instances to classify
6. Classify each instance with classification model learned by each classifier of ensemble
7. Maintain the track of each learning model or classifier for expiration situation as separate thread
8. If that instance is classified by majority of classifier or threshold based classifiers then
   a. Assign class label to that instance
   b. Send that instance to secondary memory
9. Otherwise
   a. Mark current instance as unclassified and send it to drift detection module

Algorithm-2: Drift Detection Module
Input: Current Instance or set of unclassified instances
Output: Classified instance or Send of “LearnNewModel()”
1. Initialize Recurrence Concept Drift Detection Process
2. Place oldest classifier or model of current ensemble of classifier into disk for some time
3. Read a single auxiliary classifier or model from the disk and place it into the memory by comparing of period of timestamps of auxiliary classifiers and current instance.
4. If month or time of(classifier) is approximately equal to current instance’s timestamp then
   a. Read that learning model from disk and place it into memory
   b. Classify the current instance against it
   c. If the instance is classified send it to the disk
   d. Otherwise
      i. Identify the drift using standard drift detection techniques such as EDDM
      ii. If drift is detected then send current instance to “LearnNewModel()” Module
Algorithm-3: Learn New Model

**Input:** New set of training instances, Ensemble parameters if any

**Output:** New Learning Model

1. Read current set of instances
2. Build Classifier by identify the classes
3. Apply the same approach for all other classifiers in the ensemble
4. Append the new classifier to current set of classifier in the memory

Algorithm-4: Learning Model Watchdog

**Input:** current Learning model, classification error threshold

1. Read each classifier or learning model periodically
2. Evaluate classification accuracy and performance
3. Mark the classifier as invalid if its classification error rate is above specified error threshold
4. Write marked classifiers or learning models to disk with few samples for future reference in case of recurrence concept drift.
5. Remove the marked classifier from memory.

IV. DESCRIPTION OF ALGORITHMS

Algorithm-1 is used to read the data stream in a adaptive sliding window model. In this algorithm every instance is assigned to a timestamp. This is heart of our proposed approach. This timestamp is used to search related auxiliary learning model with in short period of time. Initially none of the classifier in the ensemble is trained. So, whenever data stream is started, a new sliding window is also started. Every sliding window can read data instances from stream for some fixed predefined interval of time. First sliding window is used to train the learning model. There after every new window of points will be treated as test instances.

It is also having mechanism to classify the instances. Every instance is submitted to all the classifiers as test instance. Each classifier classifies that instance based on the underlying learning model. Based on the majority voting strategy, that instance is assigned to highest majority classification label. If that instance is not at all classified by any classifier then it is treated as outlier and kept unclassified and forward to Drift Detection Module.

Algorithm-2 In this algorithm unclassified instances are verified against learning model those are already invalidated due to concept drift. This is required to check whether that instance is any recurrence concept. In this point, it is not worthy to check all the auxiliary learning models. So we add some intelligence to the algorithm to reduce the classification time complexity. In general any recurrence relation must have relation with time or season. So very the same time region in the past learning models and try to compare the nature of the properties with those learning models only. For this purpose all the learning models are indexed internally based on timestamp. In this case time to identify right learning model is reduced and unrelated learning or less priority models are pruned in the classification process.

Algorithm-3 If no auxiliary model classifier is there to classify new instance then it means concept drift is occurred. So it is the time to build or learn new model that support this type of classification from this point onwards. So this instance is submitted to new learning model. For some time period current window of instances are used for training the new set of learning model. This new learning model will be appended to current available classifiers’ list. From that point onwards new learning model is also part of classification process. Influence of that classifier or learning model will be continued for some time. This is a long term process till the end of the stream.

Algorithm-4 This algorithm is used to monitor the classifiers. If any of the classifier is showing poor performance in its classification process then it will be invalidated by this watch dog mechanism. Later those learning models will send to persisting state for recurrence drift detection process.

V. SIMULATION RESULTS

Figure 3 the accuracy being displayed when naive's Bayesian and decision tree combination used for the classification
VI. CONCLUSION AND FUTURE WORK

Main focus of this paper is to handle recurrence drift using timestamp. To achieve this task four dependent algorithms are proposed. Finally simulation results show the drift occurrence pattern, time taken to classify the stream, kappa statistics and classification accuracy. Current work can be further extended to classify image datasets, document classification and pattern recognition.

REFERENCES


[10] Samuel Odei Danso,” An Exploration of Classification Prediction Techniques in Data Mining: The insurance domain”