Efficient Classification of Concept Drift in Data Stream

Ei Thwe Khaing
University of Information Technology
eithwekhaing@uit.edu.mm

Abstract

The classification in data streams is widely studied in the literature over the last decade. In recent literature, many research contributions use incremental or progressive learning strategies to classify the data streams. Stream classification is a variant of incremental learning of classifiers that has to satisfy requirements specific for massive streams of data. There are many methods such as single classifiers, windowing techniques, drift detectors and ensemble methods. The classifier ensembles provide a way of adapting to changes by modifying ensemble components or their aggregation method. Adaptive *Classifier Ensemble (ACE) method use to provide a natural way of adapting to change by modifying ensemble members. This method improves more accuracy and adaptable than other ensemble methods.*

Keywords – Data Stream Mining, Classification, Concept drift

1. Introduction

A data stream is a massive, infinite, temporally ordered, continuous and rapid sequence of data elements. Research on data stream is motivated by emerging applications involving massive data sets such as customer click streams, supermarket, telephone records, stock market, meteorological research, multimedia data, scientific and engineering experiments and sensor data. A new generation of mining algorithms are needed for real time analysis and query response for these applications since most conventional data mining algorithms can only be applied to static data sets that may be updated periodically in large chunks, but not in continuous streams of data [1].

While data mining has become a fairly well established field now, the data stream problem poses a number of unique challenges which are not easily solved by traditional data mining methods. Some of issues of data stream like dynamic nature, infinite size and high speed, unbounded memory requirements, Lack of global view, handling the continuous flow of data impose a great challenge for the researchers dealing with streaming data. Unlike traditional data sets, it is impossible to store an entire data stream or to scan through it multiple times due to its tremendous volume. New concepts may keep on evolving in data streams at different times, to deal with this any data stream processing algorithm have to continuously update their models to adapt to the changes.

When data streams are large, a new computational requirement needs a limited amount of memory and a short processing time in processing the continuous incoming data stream. Data stream situations pose several further requirements on learning algorithms as compared with learning from static data. It is usually impossible to store all the data from streams and only a small part of the data can be stored and used for computations within a limited time span. In most cases, the arrival speed of the incoming instances from data streams enforce their processing in the real time. The data stream mining should have the following properties: high accuracy, fast adaptation to change and low computation cost in both space and time dimensions.

Concept drift is the distribution generating the items of a data stream that changes over time. The concept drift is assumed to be unpredictable, periodic seasonality is usually not considered as a concept drift problem. If seasonality is not known certainty, it might be regarded as a concept drift problem. The core assumption, when dealing with the concept drift problem, is uncertainty about the future that the source of the target instance is not known with certainty. So, the Adaptive Classifier Ensemble (ACE) method in classifier ensembles provide concepts as adapting to changes by modifying ensemble.

The challenging issues of learning ensembles concept drift from data streams. Many classifier ensembles are typically the most often applied approaches in data stream analysis. Traditional learning methods (neural networks, Naive Bayes, nearest neighbor methods, and decision rules) are only able to process data sequentially, but do not adapt, can be easily modified to react to change.

In the further section of this article, Section 2 provides the related works to classify streaming data. Then, Section 3 provides a brief description of the data stream mining technologies. Next, Section 4 describes proposed system in streaming data. Finally, Section 5 summarizes this paper.

2. Related Works

In case of data streams, the amount of distinct options or things that exist would be massive so this

makes even the more number of cache memory or system memory are not appropriate for storing the whole stream data. The major drawbacks of data streams are the speed. Speed of information stream arrival is relatively higher than the speed of information store and process.

The paper [2] mentioned mining data streams with concept drift. The goal of the paper is to propose and validate a new approach to mining data streams with concept-drift using the ensemble classifier constructed from the one-class base classifiers. The base classifiers of the proposed ensemble are induced from incoming chunks of the data stream. Each chunk consists of prototypes and can be updated using instance selection technique when a new data have arrived. When a new data chunk is formed, ensemble model is also updated on the basis of weights assigned to each one-class classifier. The prototype selection is a promising research direction when looking for effective stream mining tools. Other research will also focus on studying influence of the size of both - ensemble model and data chunk on accuracy of the ensemble classifier.

The paper [3] described non-stationary characteristics of streaming data, prediction models are often also required to adapt to concept drifts. It surveys research on ensembles for data stream classification as well as regression tasks. Besides presenting a comprehensive spectrum of ensemble approaches for data streams, we also discuss advanced learning concepts such as imbalanced data streams, novelty detection, active and semi- supervised learning, complex data representations and structured outputs.

The paper [4] proposed a new approach in weighting ensemble components applied to stream data classification. Two theorems were presented that give the foundations for this approach. The user confidence define in these theorems. If the additional components are large, we will obtain an increase of accuracy of the whole ensemble for the entire infinite data stream. This approach is based on the observation that probability of the correct tree outcome is different in various tree sections. It achieves increasing the accuracy of the whole ensemble.

3. Data Stream Mining

Data Stream mining is the process of extracting knowledge structures from such continuous, rapid data records. Mining data streams raises new problems for the data mining community about how to mine continuous high-speed data items that you can only have one look at. Due to this reason, traditional data mining approach needs to be changed and to discover knowledge or patterns from data streams , it is necessary to develop single-scan, on-line , multilevel , multi-dimensional stream processing and analysis methods. Various procedures for extraction of information from data streams were proposed in concern with data mining.

A. Clustering

Envision an enormous measure of dynamic stream information. Numerous applications require the computerized clustering of such information into segments depending on their alikeness. In spite of the fact that there are numerous effective grouping algorithms for static information sets, grouping or dividing data streams puts extra imperatives on such calculations, as any information stream model obliges algorithms to make a single pass over the information, with limited memory and constrained calculation time.

B. Classification

There are many strategies to classify static information. This is a two stage process comprising of model development from preparing data and arrangement where the model is utilized to foresee the class names of tuples from new information sets. In a conventional setting, the training information dwell in a generally static database so scanning can be carried out many times, yet in stream information, the information stream is fast to the point that capacity to store them and scanning it several times is infeasible. Another characteristic is time varying in data streams, instead of conventional database frameworks, where just the present state is stored. This change in the nature of the data takes the form of changes in the objective classification model after some time and is alluded to as concept drift. It is a vital thought when managing stream data.

C. Association

There are two stages in algorithms for the association rule. The initial step is to find continuous item sets. All continuous item sets meet the threshold value are found and the second step is to infer association rules. In this progression, in light of the continuous item sets found in the initial step, the rules that meet the certainty basis are inferred. Nevertheless, customary association standard mining calculations are produced to take a shot at static information and, along these lines, can't be connected straight forwardly to mine association rules in stream information. New researches are directed on the most proficient method to get frequently occurring elements, association rules and various patterns in the environment of stream of data.

In these methods, this paper proposes data stream mining in classification. Data Stream Classification is a traditional supervised machine learning task. Both tasks are concerned with the problem of predicting a nominal value of an unlabeled instance represented by a vector of characteristics. The main difference between these tasks is that, in streaming

scenarios, instances are not readily available to the classifier as being part of a large static dataset, and, instead, instances are provided sequentially and rapidly over time as a continuous data stream. Therefore, a data stream classifier must be ready to deal with a great number of instances, such that each instance can only be inspected once or stored for only a short period of time.

4. Proposed System

Concept drift is the distribution generating the items of a data stream that changes over time. The concept drift is assumed to be unpredictable, periodic seasonality is usually not considered as a concept drift problem. If seasonality is not known certainty, it might be regarded as a concept drift problem. The core assumption, when dealing with the concept drift problem, is uncertainty about the future that the source of the target instance is not known with certainty. It can be assumed, estimated, or predicted, but there is no certainty.

The challenging issues of learning ensembles concept drift from data streams. Many classifier ensembles are typically the most often applied approaches in data stream analysis. Traditional learning methods (neural networks, Naive Bayes, nearest neighbor methods, and decision rules) are only able to process data sequentially, but do not adapt, can be easily modified to react to change.

A computational effective algorithm needs to be adapting of concept drift in non-stationary data stream. This paper presents a new algorithm, which adapts very quickly to concept drifts, and has been specifically designed to deal with concepts. We compare our new algorithm with various well-known learning algorithms, taking into data streaming datasets from UCI Machine Learning Repository.

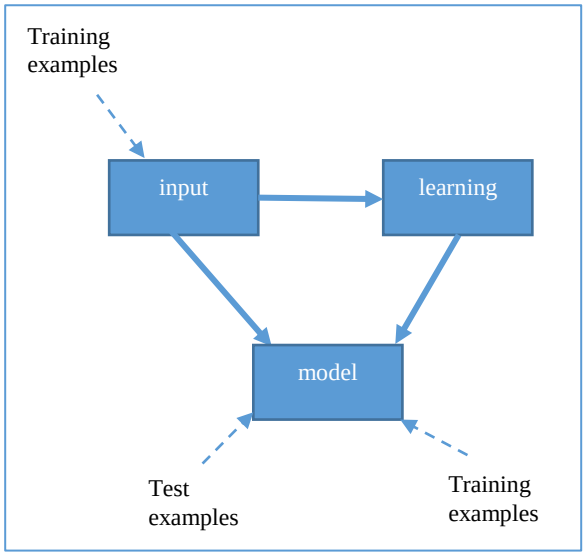

Figure 1. Overview of Data Stream Classification

5. Conclusion

The concept drifts are often unstable and change over time. The underlying data distribution may change as well. Often these changes make the model built on old data inconsistent with the new data and an updating of the model is necessary.

References

[1] R. Kalaivani, Dr. S.Vijayarani. Data Stream Mining – A Survey. IJIRCCE, Vol. 5, Issue 4, April 2017, DOI:10.15680/IJIRCCE.2017. 0504172.

[2] Ireneusz Czarnowski, Piotr Jedrzejowicz, Ensemble classifier for mining data streams, Procedia Computer Science 35 (2014) 397 – 406.

[3] Bartosz Krawczyk, Leandro L. Minku, Joao Gama, Jerzy Stefanowski, MichalWozniak. Ensemble learning for data stream analysis: A survey. Information Fusion 37 (2017) 132–156.

[4] Lena Pietruczuk, Leszek Rutkowski, Maciej Jaworski, Piotr Duda, How to Adjust an Ensemble Size in Stream Data Mining?, *Information Sciences* (2016), doi:10.1016/j.ins.2016.10.028.