Review Paper on Concept Drifting Data Stream Mining

[1] Mr. Sudhir Ramrao Rangari, [2] Ms. Snehlata S Dongre, [3] Dr. Latesh Malik

[1,2,3] M Tech (CSE), Department of Computer Science and Engineering
G. H. Raisoni College of Engineering
Nagpur, India

Abstract

The Data Stream in dynamic and emerging environment such as e-commerce, financial data analysis, sensor systems, social networking and many more fields, that possess distribution. The term concepts refer to the whole distribution of the problem in a certain point in time and hence the concept drift represents a change in distribution of the problem. Data Stream that constantly changes with time due to some hidden concepts that exhibit varying degree of drift, often the magnitude and the frequency of drifting concept are not known apriori, which is very difficult to handle, because of inadequacy of traditional techniques. However, with the advent of streaming data and long life classification system, it becomes clear that training an accurate, fast and light classifier for unpredictable, large and growing environment is very important and still open research problem.

Keywords: *Stream Data Mining, Classifier, Stream Data, Concept Drift.*

1. Introduction

In the real world where concepts are often not stable but change with time like high speed, dynamic data is generated continuously in the field of emerging environment such as telecommunication, social networking, web mining, scientific data, financial data and many more application called as data stream. A data stream is an ordered sequence of instances that show evidence of varying degree of changes. Stream Data mining is a difficult process as it dealing with data arrives in the form of continuous, high speed, large, and time varying streams and processing of such streams involve real time constraint because the speed of arrival of data is very fast. Therefore, only a small summary can be considered and each component has to be processed effectively in real time, and then discarded. The problem of data stream classification has been widely studied over last decade, the dynamic and evolving nature of data stream get attention towards research in this field. Researcher find two most challenging characteristics of data stream that are infinite length and concept drift. The concept drift occurs in the data stream when the underlying concepts of the stream changes over a time. However, online environments are often non stationary and the variables to be predicted by the learning machine may change with time. For example, the users may change their subjects of interest with time in an information system filtering, for that the learning machines should be able to model changes and adapting these environments quickly and accurately.

Several approaches have been discussed which was proposed for data stream mining over the last few years [1], [2], and [3], such as Flora Framework By Widmer G in which Window Adjustment Heuristic Method is used which is based on the matching conditions between the described items and the sample, in this method, the concept description items in training set can be divided into three categories positive Descriptor Set, Negative Descriptor set and uncertain description set. Benefit of using this, it maintain Dynamic Window to keep the track of occurrence of Drift but it deals with only one sample each time, so it has limitation on the speed of arriving data. In addition, several ensemble classifier method such as streaming ensemble algorithm and dynamic weighted ensemble have been developed so far. However, with the advent of data stream and long life classification system, it has become clear that these assumptions no longer hold, and hence training an accurate, fast and light classifier for unpredictable, large and growing environment is very important and still open research problem, Therefore There is need of a new advancement in this field which have a capability to handle concept drift in data stream quickly and accurately.

The paper is organized as, Theoretical foundation including data stream background and types of concept drift are discussed in Section 2. Section 3 discusses the review of the several methodologies that handle concept drift in data stream with conclusion in Section 4.

2. Theoretical Foundation

Data stream: This creates several challenges on data mining algorithm design. One is that the algorithms must be capable of using limited resources, time and memory, by necessity they must deal with data whose nature or distribution changes over time. In turn, dealing with time changing data requires strategies for detecting and quantifying change. A data stream is an ordered sequence of instances. Stream mining is a difficult process as it dealing with data arrives in the form of continuous, high speed, and time varying data streams, and the processing of such streams needs a real time constraint. Several techniques such as sampling, load shading and windowing have been applied in the field of data stream mining that are discussed the data stream mining and the importance of its applications, in Zaslavsky et al [4], the techniques have their roots in statistics and theoretical computer science basic. There are two category of stream mining algorithm that are Data based and task based techniques. Based on these two categories, a number of classification, clustering, time series, and frequency counting analysis have been developed.

Concept Drift: The fundamental problem in learning drifting concepts is how to identify those data in the training set in a timely manner that are no longer consistent with the current concepts and hence several criteria is used to measure the concept drift such as speed.

Speed: It is the inverse of the time taken for old concept is completely replaced by new concept. There are two forms exist, that are gradual and abrupt.

Gradual Concept Drift: The time step is taken slowly or gradually for old concept is completely replaced by new concept.

Abrupt Concept Drift: The time step is taken immediately or suddenly for old concept is completely replaced by new concept

Figure 1 shows two major types of concept change that are abrupt (sudden) and gradual.

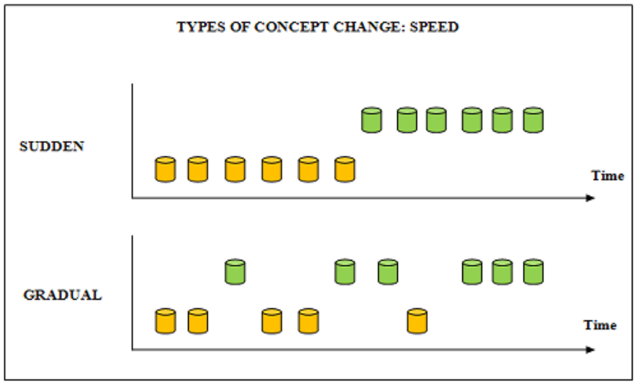

Fig 1: Types of concept change in stream data

Concept drifting data stream is difficult to analyze when it is work with real world data sets, because the magnitude and the frequency of drifting concept, type of drift and when it is started are not known a priori, even also not known there really a drift is present or not. So it is not possible to analyze detailed behavior of method or algorithm in presence of concept drift using only real world data sets. The Synthetic Data sets may further be generated by algorithms for the purpose of testing that are discussed how synthetic data is generated [13].

3. Methodology that handle Concept Drift

Data streams have gained ground attention in the field of research. The research in this field is mainly focus on the areas like query processing, and mining data streams. For instance several research have been done now in the field of data stream mining which includes the methods like classification and clustering but mining concept drifting data stream is still open research problem, So several method have been reviewed here which can handle concept drifting data stream.

Wang et al. [5] proposed a general framework for mining concept drifting data streams. This was the framework which worked on drifting. They have used weighted classifier to mine streams in which old data expires based on distribution of data. The proposed algorithm combines multiple classifiers weighted by their expected prediction accuracy. Domingos et al. [6] have developed Vary Fast Decision Tree (VFDT) which is a decision tree constructed on Hoeffding trees. The split point is found by using Hoeffding bound which satisfies the statistical measure. Haixia Chen et al [7] proposed Classifier in the ensemble is selected for integration on Hypothesis Test (CSHT) which is a new ensemble learning environment for supervised learning for concept drifting data stream. The system is based on detecting concept drift and adapting the system by classifier selection. The techniques aim to identify the usability of base classifier that representing same or similar concept with current one to improve accuracy. This approach used Naves Bayes as base classifier. Hypothesis test is used to monitor the stability of the distribution underlying the data batches over the extended period of time. Benefits of using this method is adapt the different kinds of drift and could achieve better performance but this approach does not deal with abruptly changing and conflicting concept. Confidence Distribution Batch Detection is proposed by Patrick Lindstrom el al [8]. In this, Support vector Machine, Kullback-Leiber diversion techniques are used. CDBD is a concept drift handling approach which explicitly detects changes in the data without using labeled data. In this, the classifier built from initial training data, classifies the instances in the stream and store the output of classifier in the batch. The detection algorithm calculates the indicator value on current batch and flag is changes in concept. Benefits of using this approach is that it has explicit detection mechanism which is used to detect concept change, so do not require labeled instance to detect concept drift but rebuild policy is less suitable. Teo Susnjak et al. [9] proposed a layered approach in which individual classifier combined into ensemble cluster and assigned competence weight based on performance on training data in each layer. During run time all ensemble cluster produce a collective value. This value compared with threshold to formulate final classifier, Benefits of using this approach, it can handle different types of drift such as gradual and reoccurrences but there is difficulty in determining responsiveness to concept drift between gradual adaption and performance degradation due to layer threshold update.

The adaptive ensemble boosting Classifier (AEBC) approach is proposed by K Wankhede et al [10] uses adaptive sliding window and Hoeffding Tree with naïve bayes as base learner. The Adaptive Ensemble Boosting Classifier method achieves distinct features such as; it is dynamically adaptive, uses less memory and processes data fast. The sliding window is used for change detection, time and memory management. In this algorithm sliding window is parameter and assumption free in the sense that it automatically detects and adapts to the current rate of change. If change detect, it raises change alarm. This can adapt gradual concept drift but it cannot adapt abrupt drift. WEAP-I [11] proposed by Tao Wang, is a weighted ensemble on averaging probability integrated. This is based on weighted ensemble and averaging probability ensemble under the learnable assumption. The method that which train a weighted ensemble on most n data chunks and trains an averaging probability ensemble on most recent data chunk. This approach can solve the problem of continuous concept drift occurrences. It is more robust to the averaging probability ensemble. Learn++.NSE proposed by Ryan Elwell et al. [12] allows the algorithm to identify, and perform accordingly to the changes in data distributions, as well as to recognize a possible reoccurrence of an earlier distribution. Learn++.NSE is an ensemble-based batch learning algorithm that uses weighted majority of voting, and the weights are dynamically updated with respect to the classifiers time adjusted errors on current and past environments. It employs a passive drift detection mechanism, and uses only current data for training. It can handle a variety of non stationary environments, including drift that is slow or fast, gradual, cyclical or even variable rate drift. It is also one of the few algorithms that can handle concept addition new class or deletion of an existing class but it does not focus on statistical analysis for possible performance guarantees on different NSE scenarios.

4. Conclusion

Mining concept drifting data streams is a challenging research. In particular, this paper incorporated with the review of data stream, types of concept drift and methodologies. Several important approaches that are developed so far, handle gradual concept drift in data stream but not abrupt concept drift enough. In addition, analyzing real word data by using this approach is very difficult and hence need ground attention toward this problem.

References

[1] Mahnoosh Kholghi, Hamed Hassanzadeh, Mohammad Reza Keyvanpour, “Classification and Evaluation of Data Mining Techniques for Data Stream Requirements”, International Symposium on Computer, Communication, Control and Automatio, 2010.

[2] OUYANG Zhenzhen, “Study on the Classification of Data Streams with concept Drift”, Eighth International Conference on Fuzzy Systems and Knowledge Discovery (FSKD), 2011

[3] C. Agrawal, J. Han, J. Wang, P. Yu, “A Framework for On-Demand Classification of Evolving Data Streams”, IEEE Transactions on Knowledge and Data Engineering, Volume 18(5), pp 577-589, 2006.

[4] Mohamed Medhat Gaber, Arkady Zaslavsky and Shonali Krishnaswamy, “Mining Data Streams: A Review” SIGMOD Record, Vol. 34, No. 2, June 2005

[5] H. Wang, W. Fan, P. Yu and 1. Han, “Mining Concept-Drifting Data Streams using Ensemble Classifiers”, in the 9th ACM International Conference on Knowledge Discovery and Data Mining (SIGKDD), Washington DC, USA, Aug. 2003

[6] P. Domingos and G. Hulten. “Mining High-Speed Data Streams”, In Proceedings of the association for Computing Machinery Sixth International Conference on Knowledge Discovery and Data Mining, 2000.

[7] Haixia Chen, Shengxian Ma, Kai Jiang “Detecting and Adapting to Drifting Concepts”, 9th International Conference on Fuzzy Systems and Knowledge Discovery (FSKD 2012)

[8] Patrick Lindstrom, Brian Mac Namee, Sarah Jane Delany, “Drift Detection using Uncertainty Distribution Divergence”, 11th IEEE International Conference on Data Mining Workshop(2011)

[9] Teo Susnjak, Andre L. C. Barczak, “Adaptive cascade of boosted ensembles for face detection in concept drift”, Springer-Verlag London Limited 2011

[10] Kapil K. Wankhade, Snehlata S. Dongre, Kalpana A. Mankar Prashant K. Adakane, “A New Adaptive Ensemble Boosting Classifier for Concept Drifting Stream Data”, 3rd International Conference on Computer Modeling and Simulation (ICCMS 2011)

[11] Zhenzheng Ouyang, Min Zhou, Tao Wang, Quanyuan Wu, “Mining Concept Drifting and Noisy Data Stream using Ensemble Classifier” International Conference on Artificial intelligence and Computational Intelligence(2009)

[12] Ryan Elwell, Member, IEEE, and Robi Polikar, Senior Member, IEEE, “Incremental Learning of Concept Drift in Nonstationary Environments”, IEEE TRANSACTIONS ON NEURAL NETWORKS, VOL. 22, NO. 10, OCTOBER 2011

[13] Wei Fan, “Systematic Data Selection to Mine Concept Drifting Data Streams”, KDD’04, August 22-25, 2004, Seattle, Washington USA

[14] Hanady Abdulsalam, David B. Skillicorn, Member, IEEE, Patrick Martin, Member, IEEE Computer Society, ”Classification Using Streaming Random Forests”, IEEE TRANSACTIONS ON KNOWLEDGE AND DATA ENGINEERING, VOL. 23, NO. 1, JANUARY 2011

[15] DharaniK, Kalpana Gudikandula. "Actionable Knowledge Discovery using Multi-Step Mining." International Journal of Computer Science 1.

[16] Leandro L. Minku, Student Member, IEEE, Allan P. White, and Xin Yao, Fellow IEEE, ”The Impact of Diversity on Online Ensemble learning on concept drift”, IEEE TRANSACTIONS ON KNOWLEDGE AND DATA ENGINEERING, VOL. 22, NO. 5, MAY 2010.

[17] Mohammad M. Masud, Qing Chen, Latifur Khan, Charu Aggarwal Jing Gao, Jiawei Han and Bhavani Thuraisingham, “Addressing Concept-Evolution in Concept-Drifting Data Streams”, IEEE International Conference on Data Mining, 2010

Mr. Sudhir Rangari received Bachelor of Technology in Information Technology from Swami Ramanand Teerth Marathwada University, Nanded, India in 2008. He is pursuing M. Tech degree in Computer science and Engineering from G.H. Raisoni College of Engineering, Nagpur, India. His research area is Data Stream Mining and Machine Learning.

Ms. Snehlata S. Dongre received B.E degree in Computer science and Engineering from Pt. Ravishankar Shukla University, Raipur, India in 2007 and M.Tech degree in Computer Engineering from university of Pune, Pune ,India in 2010. She is currently working as Assistant professor in the Department of Computer science and Engineering at G.H. Raisoni College of Engineering, Nagpur, India. Number of publication is in IEEE and Journals. Her research is on Data Stream Mining, Machine Learning, Decision Support system, ANN and Embedded System. Her book has published on title Data Stream Mining: Classification and Application, LAP publication House, Germany, 2010. Ms. Snehlata S. Dongre is a member of IACIST, IEEE and ISTE Organization.

Dr. Latesh Malik received B.E degree in Computer science and Engineering from Rajastan University, India in 1996 and M.Tech degree in Computer science and Engineering from Banasthali Vidyapith, India in 2001. She is currently working as Professor and head of Department in Department of Computer science and Engineering at G.H. Raisoni College of Engineering, Nagpur, India. Number of publication is in IEEE and Journals. Her research is on Document Processing, Soft Computing and Image Processing. She received Best teacher award and she is Gold medalist in M.Tech and B.E. she also received the grant from AICTE of Rs. 7.5 lacs under RPS. Dr. Latesh Malik is a member of CSI, IEEE, and ISTE Organization