CHALLENGES IN DATA STREAM CLASSIFICATION

[1]Jagannath E. Nalavade , [2]Dr. T. Senthil Murugan
[1]Research Scholar ,Veltech Dr. RR & Dr. SR Technical University, Avadi, Chennai,
[2]Department of Computer Science Engineering Veltech Dr. RR & Dr. SR Technical University, Avadi, Chennai, India
E-mail: [1]jen.sit@sinhgad.edu, [2]senthilmuruganme@gmail.com

Abstract— Stream data mining is emerging field in the data mining. Stream data generates from many applications such as banking, sensor networks, blogs at twitter. It is conceptually endless sequences of data records. It is often arriving at high rates. Analyzing or mining data streams raises several new issues compared to standard data mining algorithms. Standard data mining algorithms assume that records can be examined multiple times. Data stream mining algorithms, on the other hand, are more challenging to design since they must be able to extract all necessary information from records with only one pass over the data. Data stream mining algorithms must be online. Arrival rates for records are high, so the practical complexity of processing must also be low. Since the classification algorithm executes endlessly, it must be able to adapt the classification model to changes in the data stream, in particular to changes in the boundaries between classes ("concept drift").

To maximize usefulness, classification should be possible as soon as a sufficiently robust model has been built. The model should be updated because of a change in the underlying data. Hence, the classification algorithm should be incremental, so that changes require a model update rather than a completely new model. In the area of stream data classification many algorithms are designed but again there are certain limitations of every algorithm. In this paper, review of previous designed algorithm and challenges in the stream data classification is discussed in detail.

Keywords—Stream data classification, ensemble learning, spatial indexing, concept drifting

I. INTRODUCTION

STREAM data mining is emerging field in the data mining. Stream data generates from many applications such as banking, sensor networks, blogs at twitter. It is conceptually endless sequences of data records. It is often arriving at high rates. Analyzing or mining data streams raises several new issues compared to standard data mining algorithms.

Standard data mining algorithms assume that records can be examined multiple times. Data stream mining algorithms, on the other hand, are more challenging to design since they must be able to extract all necessary information from records with only one pass over the data.

Data stream mining algorithms must be online. Arrival rates for records are high, so the practical complexity of processing must also be low. Since the classification algorithm executes endlessly, it must be able to adapt the classification model to changes in the data stream, in particular to changes in the boundaries between classes ("concept drift").

To maximize usefulness, classification should be possible as soon as a sufficiently robust model has been built. The model should be updated because of a change in the

underlying data. Hence, the classification algorithm should be incremental, so that changes require a model update rather than a completely new model.

II. STREAM DATA MINING

Data stream continuously arrives at a system and also changing sequence of data, it is difficult to store and process. Imagine a satellite mounted remote sensor that is constantly generating data. The data is massive, fast changing, and potentially infinite. These characteristics of stream data creates challenging tasks in stream data mining. Traditional OLAP and data mining methods typically require multiple scans of data and are therefore infeasible for stream data applications. Data streams generate in many applications hence it is difficult to use or modify previous mining algorithms to fit data streams. Data stream mining has many applications and it is a hot research area. With recent progress in hardware and software technologies, different performance measures can be used in different fields. These measures are feasible for steaming data. Most of the applications requires to find new patterns for evolving new concepts. These models can be used for decision making.

Data Stream mining refers to knowledge extraction as models or rules from evolving data streams. Data Streams have different challenges like storage for storing results, querying, preprocessing and analyzing.

The first research challenge is to design fast and reliable mining models for evolving data streams. That is, algorithms that only require one pass over the data and work with limited storage. The other one is created by the highly dynamic nature of data streams, whereby the stream mining algorithms need to detect changing concepts.

In next section, we are going to discuss challenges regarding the classification of data stream.

III. REVIEW OF STREAM DATA CLASSIFIERS

Peng Zhang et al.[1] have designed tree-based indexing structure. The methodology used in this paper is Novel Ensemble-tree classifier indexing structure. Advantages of this classifier is E-trees are updated automatically. It can be done by integrating new classifiers. This classifier also solves prediction efficiency problem for ensemble models. With above advantages several disadvantages present in this algorithm. Some are obtains lower error rate at a cost of heavier time cost. It also consumes a larger memory space.

Brzezinski et al.[2] have designed Hybrid algorithm, called Accuracy Updated Ensemble Algorithm. It is also called as Accuracy Updated Ensemble (AUE2). An advantage of this algorithm is reacting to different types of concept drift much better than related adaptive ensembles. It improves the ensemble's reactions to different types of concept drift as well as reduces the impact of the chunk size. Disadvantages of this algorithm are that it not works in a truly incremental fashion in partially labeled streams and adapting weight for different data space seems tough.

Karim R et al.[3] have designed An adaptive ensemble classifier for stream data classification. It efficiently handles Complex Noisy Instances in Data Streams. It gives flexible model for data streams classification. Disadvantages are it requires more time as compare to other algorithms.
Categorical data cannot be handled by using this classifier.

Parker B et al.[4] have designed Incremental Ensemble Classifier for addressing Non-Stationary Fast Data Streams. In this algorithm Modal Mixture Model(M3) uses a mixture or ensemble of base classifiers. It addresses the core tenants of Fast Data, including high velocity, data variety (mixed types, features, and distributions), and volume (limited storage). It provides incremental class label prediction. Accuracy is better as compared to other baseline approaches. Categorical data cannot be handled by using this algorithm. It cannot detect novel classes that occur as the label space evolves throughout the data stream.

Song G et al.[5] have designed New Ensemble Method for Multi-label Data Stream Classification in Non-stationary Environment. Multi-label Dynamic Ensemble (MLDE)

approach. It deals with multi-label stream Classification. The number of selected base classifiers is changed in according to whether the concept drifts or not. It not works in noisy stream environment.

Liao J et al.[6] have designed an Ensemble Learning Approach for Concept Drift. Accuracy and Growth rate updated Ensemble (AGE). It combines the advantages of single classifier and ensemble method more flexibility on reacting to various types of concept drifts. Partially labeled data problem of data stream cannot handled.

M. Masud et al.[7] have designed Classification and Adaptive Novel Class Detection of Feature-Evolving Data Streams. Novel class detection technique. It can distinguish between different classes of data and discover the emergence of a novel class. It addresses the conceptevolution problem. This technique can be applied on both synthetic and real-world data and obtain much better results. Data stream classification problem under dynamic feature sets are not addressed. Also supervision is necessary for Classification.

Liang H et al.[8] have designed a new Data Stream Classification Algorithm. It uses Novel Digital Hormone Model (DHM) based approach. It does not need a big sample buffer in the classification process and the classifier can be updated efficiently -ability to predict the class label accurately and to store temporary records with more smaller memory space. Cost is more as compared to some other algorithms. It cannot handle multiclass problem.

Godse A et al.[9] have designed Classification of Data Streams with Skewed Distribution CDSSDC (Classifier for Data streams with skewed distribution of classes). It gives good results than use of SMOTE. Degradation in performance when number of attributes are more.

Gomes H et al.[10] have designed SAE: Social Adaptive Ensemble Classifier for Data Streams. SAE classifier efficiently adapts itself in the occurrence of a concept drift. SAE is a reasonable choice for problems that contains abrupt or gradual drifts and in which processing time is a concern. If there are no drifts, then SAE does not perform much better than a single Hoeffding Tree.

Abdulsalam H et al.[11] have designed Classification Using Streaming Random Forests. Streaming Random Forests Algorithm achieves high classification accuracy. Random Forests is extended and name given as Streaming Random Forests. It Handles Concept drift problem easily. Categorical data cannot be handled. This algorithm not addresses missing data problem.

M. Masud et al.[12] have designed classifier to handle Class Detection in Concept-Drifting. It handles Data Streams by considering Time Constraints. Novel class detection technique can differentiates between different classes of data. As well as it discovers the occurrence of new novel class. It addresses the concept-evolution problem. It uses time constraints for delayed data labeling and making classification decision. This technique can be applied on both synthetic and real-world data and can obtain much better Results. In this algorithm data stream classification problem under dynamic feature sets are not addressed. Supervision is necessary for Classification.

M. Masud et al.[13] have designed Classification and Novel Class Detection in Data Streams with Active Mining. ACT Miner and Active classifier work with K-NN and decision tree. It Works on the less label instance. It saves 90% or more labeling time and cost. It is not directly applicable to multiclass. Also, it not works for the multi label classification.

Law Y et al.[14] have designed an Adaptive Nearest Neighbor Classification Algorithm for Data Streams. It is Incremental classification dynamic update. Storage memory problem is disadvantages of this algorithm. It is time consuming and costly learning.

Chi Y et al.[15] have designed Loadstar: A Load Shedding Scheme for Classifying Data Streams. It uses decision Trees. High speed, Need less memory space are advantages of this algorithm. Disadvantages are non-adaption to concept drift. It is time consuming and costly learning.

IV. REQUIREMENTS

In this section we discussed several requirements of stream data classification algorithm. A classification algorithm must

address following requirements for learning from evolving data streams.

Requirement 1: Process continuous arriving records in a single pass.

The main feature of a data stream is that data arrives continuously one record after another. Random access of this data is restricted. Every record must be accepted as it comes in the order. Once it is analyzed, it will not be available or we cannot access again after some time. There is no rule for preventing classifier model from storing records internally in the short term. But it creates new problem which is discussed next.

Requirement 2: Storage requirement

The main challenge is processing of data that is many times larger than the available storage capacity. Processing of such type of data is not easy. Storage capacity used by an algorithm can be divided into two types. One is memory used to store current statistics, and memory used to store the classifier model. Current statistics should directly constitute the classifier model used for classification.

Requirement 3: Prediction within very less amount of time Designed classifier's runtime complexity should scale in linear fashion for arriving unlabelled streaming data. Means prediction time should not increase as data arrives in undefined manner. Classifier must be capable of working in real time manner. Streaming data may be arrives uncertainly, classifier must process as it arrives. Algorithms fails to follow this requirement means it will loss streaming data. When data source is persistent, then absolute timing is not a critical. The slower the algorithm has the less value according to user's point of view, since they expects results in a reasonable amount of time.

Requirement 4: Always be ready to predict streaming data with new concepts at any point.

An ideal classifier should always ready to classify arriving uncertain data at any point. Once model has been built, it should classify unlabelled data as and when it arrives. Algorithm should be capable of producing the ideal model from the data it has observed. During some time periods model stays constant, for example when model is based on a batch based algorithm. It takes some time for storing up the next batch. The process of generating the model should be as efficient as possible, the best case being that no translation is necessary. That is, the final model is directly updated during running time as new concepts arrive. It should recompute the model based on running statistics.

V. RESEARCH ISSUES

The following are some of the research issues in data Streams classification technique:

- Low memory requirements while processing large data sets
- Variants in mining tasks those are desirable in data streams
- High accuracy in the results
- Transferring results over a WSN with a limited bandwidth.
- The needs of real world applications
- Efficiently handle the storage of the streaming data with timeline, as well as retrieve them during a certain time interval in response to user need.
- Algorithm built model which helps user to modify the parameters during processing period.

VI. CONCLUSION

In this paper we first discussed need of steam data mining in different applications. We even reviewed data stream classification with the existing algorithms. From the discussion we can conclude that most of the current mining approaches uses one pass mining algorithms and few of them even address the problem of drifting. Most algorithms handle data with only numerical or ordinal attributes. Real datasets contains Categorical data. It is, therefore, essential to consider data with categorical attributes.

Present algorithms not dealing with missing data. It is a common problem. This is also important challenge in the streaming data.

Many algorithms require the number of classes as an input parameter and it does not change through execution. An interesting problem for stream data classification systems, however, is that classes might split (or join) as a

result for concept drift, or even a new class is introduced to the classification problem.

Research in data streams is still in its early stage. If the problems are addressed above are resolved by developing efficient classifier, in the near future data stream mining will play an important role in the business world as the data flows continuously.

References

[1] Peng Zhang, Chuan Zhou, Peng Wang, Byron J. Gao, Xingquan Zhu, and Li Guo," E-Tree: An Efficient Indexing Structure for Ensemble Models on Data Streams", IEEE Feb-2015.

[2] Dariusz Brzezinski and Jerzy Stefanowski," Reacting to Different Types of Concept Drift: The Accuracy Updated Ensemble Algorithm", IEEE Jan-2014.

[3] Md. Rejaul Karim, and Dewan Md. Farid," An Adaptive Ensemble Classifier for Mining Complex Noisy Instances in Data Streams", IEEE Jan-2014

[4] Brandon S. Parker, Latifur Khan," Incremental Ensemble Classifier Addressing Non-Stationary Fast Data Streams", IEEE-2014 Ge Song, Yunming Ye," A New Ensemble Method for Multi-label Data Stream Classification in Non-stationary Environment", IEEE July-2014

[5] Jian-Wei Liao, Bi-Ru Dai," An Ensemble Learning Approach for Concept Drift", IEEE-2014.

[6] Mohammad M. Masud,, Qing Chen, Latifur Khan, Charu C. Aggarwal, Jiawei Han, Ashok Srivastava, and Nikunj C. Oza," Classification and Adaptive Novel Class Detection of Feature-Evolving Data Streams", IEEE July -2013

[7] Hong-shuo Liang, Li-qun Jin1, Li Zhao1," A New Data Stream Classification Algorithm", IEEE-2013

[8] Abhijeet Godase, Vahida Attar," Classification of Data Streams with Skewed Distribution", IEEE-2013

[9] Heitor Murilo Gomes and Fabrício Enembreck, " SAE: Social Adaptive Ensemble Classifier for Data Streams", IEEE-2013

[10] Hanady Abdulsalam, David B. Skillicorn, Member, IEEE, and Patrick Martin," Classification Using Streaming Random Forests", IEEE Jan-2011

[11] Mohammad M. Masud,, Qing Chen, Latifur Khan, Charu C. Aggarwal, Jiawei Han ,"Classification and Novel Class Detection in Concept-Drifting Data Streams under Time Constraints", IEEE June -2011

[12] Mohammad M. Masudy, Jing Gaoz, Latifur Khany, Jiawei Hanz, and Bhavani Thuraisingham, "Classi_cation and Novel Class Detection in Data Streams with Active Mining", Springer 2010

[13] Yan-Nei Law and Carlo Zaniolo," An Adaptive Nearest Neighbor Classification Algorithm for Data Streams", 9th European Conf on Knowledge Discovery in Databases, Springer Verlag, 2005

[14] Chi Y., Wang H., and Yu P.S. "Loadstar : Load Shedding in Data Stream Mining", in Proceedings of the 31st VLDB Conference,Trondheim, Norway. pp. 1302-1305, 2005.

Table 1. Analytical Framework of Stream Data Classification

.

Sr.No	Title	Author	Mining Task/ Methodology	Advantages	Disadvantages
1	E-Tree: An Efficient Indexing Structure for Ensemble Models on Data Streams	Peng Zhang et al.[1]	Novel Ensemble-tree classifier indexing structure	-E-trees can be automatically updated by continuously integrating new classifiers - prediction efficiency problem for data streams solved	-obtains lower error rate at a cost of heavier time cost-E-trees consume a larger memory space
2	Reacting to Different Types of Concept Drift: The Accuracy Updated Ensemble Algorithm	Brzezinski et al.[2]	Hybrid algorithm, called Accuracy Updated Ensemble Accuracy Updated Ensemble (AUE2)	-react to different types of concept drift much better than related adaptive ensembles -improve the ensemble's reactions to different types of concept drift as well as reduce the impact of the chunk size	-It not works in a truly incremental fashion in partially labeled streams -increases memory requirements and algorithm processing time.
3	An Adaptive Ensemble Classifier for Mining Complex Noisy Instances in Data Streams	Karim R et al.[3]	An adaptive ensemble classifier	-It efficiently handles Complex Noisy Instances in Data Streams -It shows great flexibility and robustness in data streams classification	-It requires more time as compare to other algorithms -Categorical data cannot be handled
4	Incremental Ensemble Classifier Addressing Non-Stationary Fast Data Streams	Parker B et al.[4]	Modal Mixture Model(M3) uses a mixture or ensemble of base classifiers	-It addresses the core tenants of Fast Data, including high velocity, data variety (mixed types, features, and distributions), and volume (limited storage)	-time consuming -Categorical data cannot be handled

				-It provides incremental/on-line label prediction with high accuracy compared to baseline approaches	
5	A New Ensemble Method for Multi-label Data Stream Classification in Nonstationary Environment	Song G et al.[5]	Multi-label Dynamic Ensemble (MLDE) approach	-It deals with multi-label stream Classification -The number of selected base classifiers is changed in according to whether the concept drifts or not	-not works in noisy stream environment
6	An Ensemble Learning Approach for Concept Drift	Liao J et al.[6]	Accuracy and Growth rate updated Ensemble (AGE)	-It combines the advantages of single classifier and ensemble method -more flexibility on reacting to various types of concept drifts	-partially labeled data problem of data stream not handled
7	Classification and Adaptive Novel Class Detection of Feature-Evolving Data Streams	M. Masud et al.[7]	Novel class detection technique	-It can distinguish between different classes of data and discover the emergence of a novel class -It addresses the conceptevolution problem -this technique can be applied on both synthetic and realworld data and obtain much better results	-not addressed data stream classification problem under dynamic feature sets - supervision is necessary for Classification
8	A New Data Stream Classification Algorithm	Liang H et al.[8]	Novel Digital Hormone Model (DHM) based approach	-It does not need a big sample buffer in the classification process and the classifier can be updated efficiently -ability to predict the class label accurately and to store temporary records with more smaller memory space	-Cost is more as compared to other algorithms -Cannot handle multiclass problem
9	Classification of Data Streams with Skewed	Godse A et al.[9]	CDSSDC (Classifier for Data streams with skewed distribution of classes)	-It gives good results than use of SMOTE	-degradation in performance when number of

	Distribution				attributes are more
10	SAE: Social Adaptive Ensemble Classifier for Data Streams	Gomes H et al.[10]	Social Adaptive Ensemble (SAE) classifier	-It efficiently adapts itself in the occurrence of a concept drift	-If there are no drifts, then SAE does not perform much better than a single Hoeffding Tree
11	Classification Using Streaming Random Forests	Abdulsalam H et al.[11]	Streaming Random Forests Algorithm	- It achieves high classification accuracy - It combines the ideas of streaming decision trees and Random Forests -It Handles Concept drift problem easily	- Categorical data cannot be handled - This algorithm not addresses missing data problem.
12	Classification and Novel Class Detection in Concept-Drifting Data Streams under Time Constraints	M. Masud et al.[12]	Novel class detection technique	-It can distinguish between different classes of data and discover the emergence of a novel class -It addresses the concept-	-In this algo data stream classification problem under dynamic feature sets are not addressed.
				evolution problem - Uses time constraints for delayed data labeling and making classification decision - Can apply our technique on both synthetic and real-world data and obtain much better Results	- Supervision is necessary for Classification
13	Classification and Novel Class Detection in Data Streams with Active Mining	M. Masud et al.[13]	ACT Miner and Active classifier work with K-NN and decision tree.	-It Works on the less label instance. -It saves 90% or more labeling time and cost.	-Not directly applicable to multiclass. -Not work for the multi label classification.

14	An Adaptive Nearest Neighbor Classification Algorithm for Data Streams	Law et Y al.[14]	Incremental classification	dynamic update	-Low speed -Storage memory problem -Time consuming and costly learning
15	Loadstar: A Load Shedding Scheme for Classifying Data Streams	Chi Y et al.[15]	Decision Trees	-High speed -Need less memory space	-Non-adaption to concept drift -Time consuming and costly learning