Data Stream Mining Developments and Applications

Jayendra Kumar, Anitha Raju

Abstract**:** ***Organizations in multiple domains create diversified data each day. Such data can be processed to improve instantaneous decision making decisions. However, it is challenging to act on real-time searching and processing large-scale datasets in an online data processing. Selection as the best technology for selecting appropriate data from a large dataset, thereby using upcoming features to make a decision. Though the number of processing data is large the accuracy of the mining counts on the number of instances in a cluster and the security measurers offered to preserve the data. The selection algorithm suggests a variety of approaches in this field to find appropriate algorithms and checks the usage used by the data streaming approach in mining concern. This paper presents a brief review on the developments and applications of data stream mining in current usage.***

Index Terms**:** ***Data stream mining, applications, security measure, Big-data research.***

I. INTRODUCTION

In recent years, the collection of knowledge from data streams has increased. However, since most current studies are concerned, to the mode of data streams it is relatively needed to optimize it. If a data stream class is unbalanced, the search complications are very difficult. In general, selected approaches adopted in the current stream mining are applied by using the existing updating methods, are getting limited used for previous mining with current usage. However, unequal assessment of cluster is affected by data streaming in mining approach. These existing techniques ignore the majority of class distribution, and hence limited to formation of appropriate clusters. The limitation of cluster formation leads to lower mining accuracy, and delay overhead in the system. The concern of data privacy in data stream mining is as well an additional overhead to the existing system. An outline to the approaches made in solving the observed issues is addressed in this paper.

II. DATA STREAM APPLICATION

The mining program in the context of data stream widely studied in the recent past [1]. With rapid progress of increasing data exchange and online learning data stream mining (DSM) has got lot of attention. Usually, data stream mining implies mining activities that makes a search of data files presented in the online stream data. When continuously exchange data is collected, data distribution called concept drift [2] is observed. It is defined as a dynamic change of data distributions and a data stream mining technique must have the capacity to create and dynamically modify mining model to overcome the concept drift. Over the past few decades, researchers have developed various approaches [3] to tolerate concept drift. Further attention was given to the involve the concerns of complexity of the classical social probability distribution observed in DSM application. Some literature is class-intensive study [4] and hypothesis [5] were observed in this category. Though methods have shown better performance, they are either limited of a usage due to complexity or with the concern of resource and privacy issue.

As class evolution is concerned with a particular situation of concern, various interfacing evolutions were observed. For example, Smartphone with the window protocol affects the complex influence of data stream and influence on the buffering in data stream application. New DSM approaches need to develop with abandoning approach, having constraint to forget factor about useful information in Mining application. Inevitably, a complex practice include memories, and live buffering of the stream data [6,7]. In the approach of performance improvement in DSM, researcher has outcome with different methods by sharing information on each base learner [8]. A different set of basic approach dealing with an expert class evolution helping in creating different classes for basic learner system is developed. For example, in terms of class indifference, the collective measures of former base information's are give higher observation than the current data [9]. Information on an on-line basis for each base learner takes time to evolve the classes. Hybrid coding methods help to combine additional programs and on-line memory with the advantage of providing a structure within the architecture such as AUE2 [10] algorithm. This algorithm uses every part of the stream for the intimation of Basic learning in DSM application. The method of initiation hangs up their precision for categorization. Basically in an open category, the data are not balanced for the current basic learners. For the first stage class evolution it gradually evolves and the class of the data is difficult to identify effectively. Beyond previous strategy, link discovery methods defined the concept of expression and modified for the cycle of discovering information from the data stream. This method has a constraint of forgetting the descending window useful information in the evolution process. DDD [3,11] is defined as a detection approach that preserves old records while using the new inputs. Nevertheless, only the old data is maintained for the previous details of the information stream. In this case of class evolution, more than one class evolutionary process takes place from time to time and the DDD forget the buffered information after a specified period.

Revised Manuscript Received on June 15, 2019.

Mr Jayendra Kumar, Research Scholar, CSE, Koneru lakshmaiah Education Foundation, Vaddeswaram, Guntur Dist,.A.P.India.Pin:522502, Email:Jayendrakumar.sahu@gmail.com.

Dr. Anitha Raju, Associate Professor, CSE, Koneru lakshmaiah Education Foundation, Vaddeswaram, Guntur Dist.,A.P.Pin:522502, Email:anitharaju@kluniversity.in.

Class evolution is relevant to the increase of concept of repeating data in a class, which refer to the previous information, for varying parameter. The recurrent details of the system as well the entire class structure that includes the same information is set as a group termed 'cluster'. While the class are aligned to cluster some classes disappear with evolution. Class rearrangement therefore leads to a repeating concept, and existing algorithms of repeated complexity cannot be effectively handled by these methods.

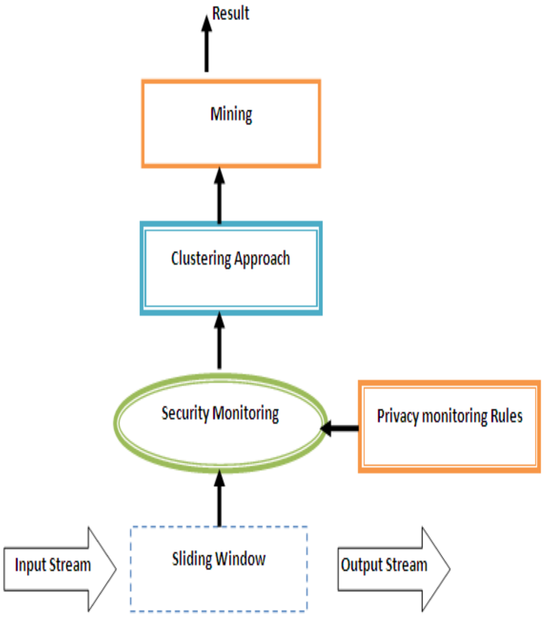

Figure. 1 Generalized flow of Data stream mining application .

In short, research progress on generalization of DSM has provided inspiration for resolving class evolution, with some of the approaches applied directly to this particular concern. So, a dedicated research on the classification of classes on cluster is more than a decade old. In clustering, [12] introduced an idea of increasing the class with evolution from the time of initiation to cluster formation. Algorithm for the class alignment for cluster is outlined in [13]. A Chunk of stream is processed by data stream mining in class evolution. The primal point of this approach is aimed for effectively retaining the model for use with monitoring the complexity of operation [14].

To characterize the core character in DSM i.e., a basic work model defined with two process.

(1) Unconfident data stream of varying modes used of training without class details. Here, the concern on the reliability of the chosen data is defined by the outline structure of the stream. The base learners mark instances of novel classes in classes across the stream [15]. This however, mislead in clustering as without the selection criterion followed. For the generality of class, the clustering approach removes obsolete data from segregation. However, once the class is defined, this model needs to be updated periodically.

(2) In clustering process, all the individual class becomes a reference in defining the data grouping. CLM-k-means [16] is an approach defined to determine the decision of a class. In this approach however, the clusters are determined based on a regular value transition of class.

A high k value is not suitable for a class to emerge or disappear, for instance when its size is larger, it may lead to a small amount of unsatisfactory performance. Other algorithms associated with class evolution include families, such as study Learn++ [17], Learn++.NC [18], which are adaptive in nature based on a set of basic detail for each patch. When forming a class based on the traditional chunk value, the former data who were trained without this class are updated as new entry. DW-CAV [19], is introduced as an approach of clustering with a dynamic weight adaptation. To find a balanced data, sub clusters are projected. [20, 21] defined a clustering Strategy operates as a warehouse. The approach has a constraint in:

(1) The basic learners is complex in approach.

(2) In the process of operating this algorithm, each basic learner needs to make sure that the total weight for value below a limiting value. If so, the dynamic irregular complexity and more than one class in data classes is not suppressive. In a gradual updation, the approach takes time to determine the weight parameter.

III. Security Measures

In data streaming publishing data need to maintain the privacy and should operate for publishing useful information on maintaining personal privacy. Recently, security in DSM has gained much attention in academics, industries, and various data publishing domain [22]. Here, Privacy implies data containing details of transaction for identification and transaction details. With the emergence of large data streaming, data can be repeated in limited and growing level to tolerate the attacks [23]. Multilingual online applications such as network analysis [24] provide frequent transmission transactional data streams for security provision. However, disclosure of the data stream affects the privacy in the network. The privacy issue for publishing transactional data has also gained some interest in recent past. However, the nature of streaming and the approaches of privacy threatening has constraint the transaction performance. In [25] transactions are given with time period known as landmarks and are considered until the current time of transaction. In the process of sliding window the input data is processed with window given for data mining called as transactions-sensitive windows and time sensitive windows. These windowing are applicable to data, as transactions lead to the removal of the transactional data for higher volume of data.

To develop a privacy measure in transaction data in [26] a Model is presented, which set most of the background scenario by reducing a k-factor for related data. Here the Anonymity mechanism has been incorporated by monitoring and cooperating to privacy coding in streaming data. In [27] Background knowledge was transferred to clients, which defines for a transaction.

In [28] an attacker is assumed to have an unregulated number of intrusions, and presented a monitoring system, which observe the groups, and then has sensitive values for each group defined to monitor.

Published By:
Blue Eyes Intelligence Engineering
& Sciences Publication

Table. 1 Summarized outlined of Comparative Performance Analysis Data stream mining.

Refren ce No	Approach	Method	Performance	Limitation
[1]	Mining high speed data stream	Decision method	4-12%	Web log data, non discretized attribute
[2]	Diversity in concept drift	Ensemble	2-10%	Recurrent drift
[3]	Evolving data stream	Diversity Dealing Drift	2-8%	Skewed data set
[4]	Mining concept drift data stream	Hybrid decision tree	2-10%	Attribute-Incremental Learning
[5]	Mining concept drift data stream	Novel class detection	3-12%	Multi label classification
[6]	Ensemble of classifier	Learn ++NC	3-18%	Non-stationary learning
[7]	Novel class detection	Time constraint	1-6%	Dynamic feature set
[8]	Stream classification	Class based ensemble	3-9%	Dynamic windowing approach
[9]	Learning from time varying data	Adaptive window	2-10%	Real world data set
[10]	Mining data stream	Ensemble	-	Speed and accuracy
[11]	Mining data stream	Ensemble	-	High dimensional data set
[12]	Mining concept drift data stream	Ambiguous decision tree	3-15%	Incremental fashion in partially labeled stream
[13]	Mining concept drift data stream	Recurring and novel class	2-9%	Monitoring the learning process
[14]	Mining concept drift data stream	Recurring and novel class	2-9%	Multi threading, parallelism
[15]	Novel class detection	Time constraint	1-6%	Drift detection
[17]	Supervised learning	Learn ++NC	3-15%	Comparison with weak learner
[18]	Ensemble of classifier	Learn ++NC	3-18%	Evaluation on Severe unbalanced data
[21]	Over sampling	Synthetic Minority Over sampling Tech	4-14%	Larger decision regions
[23]	Online transactional data stream	Varying size sliding window	3-18%	Simulation experiment
[31]	Data stream using sliding window	Max-frequent item set	2-11%	Concept drift
[32]	Real time stream data mining	CanTree, GTree	2-13%	Dynamic data
[33]	Real time stream data mining	CanTree, GTree	2-13%	Efficient data structure
[35]	Privacy preserving	Set valued data	2-12%	Low information loss
[36]	Privacy preserving	Top down, local generalization	3-17%	Local recording techniques
[37]	Privacy preserving	Top down, local generalization	3-17%	Diversity in anonymizations
[38]	Anonymization transactional data	Sparse coding	4-19%	Anonymizations of high dimensional data

Defines an anonymity approach where data is defined in easy and accessible format rather than a non-formal recording using a single and generalization hierarchy [29]. [30] Introduces a controlling approach which defines the conversion, which share transactions with the maximum knowledge of materials and can be shared with a public privacy. In addition, they define global operation to protect several possible attacks for privacy constraints [31]. In all respects, the opponent's background knowledge was confined as object of observation. However, an attacker will receive partial knowledge of sensitive items and therefore, the idea of uncertainty in the privacy model will not allow an attacker to know the strategy used in defining the confidence of any section of data.

IV. PRIVACY CODING

In Data stream operation data are persistent and ambiguous, and are usually unrelated [32]. Publishing data with security measure related methods are outlined in [33] for security provisioning in DSM. To introduce a unify data stream using k-anotomy [34] approach an aided input is defined for security application. In [35] for time constraint data publication a cluster reuse, and an approach of data streams was introduced by clustering based on the speed limit of the process with reduction in information loss. In [36] a method was designed to help diagnose a stream of information in the implementation of controls for approximately obsolete data. An application of anonymous data exchange in this process is defined. In [37] an advanced approach developed to support the flow of a repetitive data to create clusters containing data elements, and clusters satisfying the authenticity are proposed. In addition, the process limits data loss for generalization. In [38] approach a delayed form of monitoring using the collection and isolation of data stream in mining application is outlined. This creates a storage delay due to anonymizing using fake values. To summarize the developed approach for Data stream mining Table 1 outline a summarized details of methods developed in the domain of data stream mining.

V. CONCLUSION

Data stream mining is observed to have a large impact in next generation applications. Selection of processing approach relevant in machine-learning applications are needed to improve the performance of data stream mining .Because the amount of data in data stream is large, the concern of update and deletion are crucial. In the development of the data mining approach, the clustering approach has a great impact on the performance on mining. Clustering using the distance metric are predominantly been observed in this domain. The approach of security provisioning has a greater effect on the accuracy and integrity of data stream in DSM application. In the development of these approaches, the need of security measure, resource overhead and delay constrain need to be addressed to improve the performance of DSM application.

REFERENCES

1. Pedro Domingos and Geoff Hulten,"Mining high speed data streams",in Proc.6th ACM SIGKDD Int. Conf. Know. Discovery Data Mining,pp:71-80,2000.
2. Leandro L Minku, Allan P white, Xin Yao,"The impact of diversity on online ensemble learning in the presence of concept drift",IEEE Transaction knowledge data Engg, Vol:22,No:5,pp:730-742,May 2010.
3. Leandro L Minku and Xin Yao,"DDD a new ensemble approach for dealing with concept drift", IEEE Transaction knowledge data Engg, Vol:24,No:4,pp:619-633,Apr 2012.
4. Zhi-Hua Zhou and Zhao-Qian Chen,"Hybrid decision tree",Know. Based syst., Vol:15,No:8,pp:515-528,2002.
5. Mohammad M. Masud, Jing. Gao, Latifur. Khan, Jiawei Han, and Bhavani Thuraisingham,"Integrating novel class detection with classification for concept drifting data streams", in Proc Eur. Conf. Mach. Learn. Know. Discovery data base, Vol:5782,pp:79-94,2009.
6. Michael D. Muhlbaier, Apostolos Topalis, and Robi Poli, "Learn ++ NC combining ensemble of classifier with dynamically weighted consultant vote for efficient incremental learning of new classes",IEEE Transaction on Neural Network, Vol:20,No:1,pp:152-168,Jan 2009.
7. Mohammad Masud, Jing Gao, Latifur. Khan, Jaiwei Han, and Bhavani Thuraisingham, "classification and novel class detection in concept drifting data streams under time constraints ",IEEE Transaction knowledge data Engg, Vol:23,No:6,pp:859-874,Jun 2011.
8. Gerhard Widmer and Miroslav Kubat, "Learninig in the presence of concept drift and hidden context", Mach. Learn.,Vol :23,No:1,pp:69-101,1996.
9. Albert Bifet and Ricard Gavalda,"Learning from time changing data with adaptive windowing ",in Proc. SIAM Int. Conf data mining ,pp:443-448,2007.
10. W. Nick. Street and YongSeog. Kim, "A streaming ensemble algorithm (SEA) for large scale classification", in Proc.7th ACM SIGKDD Int. Conf. Know. Discovery Data Mining,pp:377-382,2001
11. Matthew Karnick, Metin Ahiskali, Michael D. Muhlbaier and Robi Polikar,"Learning concept drift in non stationary environments using an ensemble of classifier based approach", in Proc. IEEE Int. Joint Conf. on Neural Network, pp:3455-3462,June 2008.
12. Dariusz Brzezinski and Jerzy Stefanowski,"Reacting to different types of concept drift :The accuracy updated ensemble algothm", IEEE Transaction on Neural Network learning syst., Vol:25,No:1,pp:81-94,Jan 2014.
13. J João Gama and Petr Kosina, "Recurrent concepts in data stream classification",Know. Infor. Syst.,Vol:40,No:3,pp:489-507,2014.
14. Sakthithasan Sripirakas and Russel Pears,"mining recurrent concepts in data streams using the discreate fourier Trans.",in Proc. 16th Int. Conf. data warehousing Know. Discovery ,pp:439-451,2014.
15. Mohammad M. Masud, Qing Chen, Latifur Khan, Charu C. Aggarwal, Jing Gao, Jiawei Han, Ashok Srivastava, and Nikunj C. Oza, "classification and adoptive novel class detection of feature evolving data stream", IEEE Transaction knowledge data Engg, Vol:25,No:7,pp:1484-1497,Jul 2013.
16. Anil. K. Jain and Richard C. Dubes, "Algorithm for clustering data ",Upper Saddle River, NJ. USA: Prentice Hall,1988.
17. Robi Polikar, Lalita Udpa, Satish S. Udpa, and Vasant Honavar,"Learn ++ :an incremental algorithm for supervised nearal networks ", IEEE Transactions . system, Man Cybern C, Appl. Rev. Vol :31,No:4,pp:497-508,Nov 2001.
18. Gregory Ditzler, Michael D. Muhlbaier, and Robi Polikar, "Incremental learning of new classes in unbalanced datasets Lear ++ UDNC",in Proc. 9th Int. Conf. Multiple Classifier syst.,pp:33-42,2010.
19. Gregory Ditzler, Gail Rosen, and Robi Polikar,"Incremental learning of new classes from unbalanced data",in Proc. Int. joint conf. Neural Network.,pp:1-8,Aug 2018.
20. Yoav Freund and Robert E. Schapire,"A decision Theoretic generalization of online learning and an application to boosting ",in Proc. 2nd Annu. Eur. Conf. Computer Learn Theory,Vol:904,pp:23-37,1995.
21. Nitesh V. Chawla, Kevin W. Bowyer, Lawrence O. Hall, and W. Philip Kegelmeyer, "SMOTE :Synthetic minority over sampling technique",J. Artif. Intell. Res.,Vol:16,pp:321-357,Jan 2002.
22. Shuo Wang, Leandro L. Minku and Xin Yao, "online class imbalance learning and its application in fault detection",Int. J. Compute. Intell. Apple.,Vol:12,No:4,pp:1340001(19 pages),2013.

23. Jyrki Kivinen, Alexander J. Smola, and Robert C. Williamson, “online learning with kernels”,IEEE Trans. Signal process.,Vol:52,No:8,pp:2165-2176,Aug 2004.
24. Nathalie Japkowicz, “Concept learning in the presence of between class and within class imbalances”,in Proc. 14th Biennial Conf. Can. Soc. Compute. Stud. Intell. Adv. Artf. Intell.,PP:67-77,2001.
25. Taeho Jo and Nathalie Japkowicz,”class imbalances versus small disjuncts”, SIGKDD Explore. Newsl., Vol:6,No:1,pp:40-49,Jun 2004.
26. Pavan Kumar Mallapragada, Rong Jin and Anil Jain, “Non parametric mixture model for clustering ”,in Proc. Int. Conf. Struct. Syntactic and statistical pattern Recog.,Vol:6218,pp:334-343,2010.
27. Haibo He, and Edwardo A. Garcia, “Learning from imbalanced data”, IEEE Transaction knowledge data Engg, Vol:21,No:9,pp:1263-1284,Sep 2009.
28. Shuo Wang ,Leandro L. Minku , Xin Yao, ”Resampling based ensemble methods for online class imbalance learning ”, IEEE Transaction knowledge data Engg, Vol:27,No:5,pp:1356-1368,May 2015.
29. Hessam Zakerzadeh,Charu Aggarwal and Ken Barker,”Managing dimensionality in data privacy anonymization”,Know. Inf. Syst.,vol. 49, no. 1, pp. 341_373, Oct. 2016.
30. Sergio Ramirez-Gallego, Bartosz Krawczy, Salvador Garcia, Michal wozniak, Francisco Herrera,”A survey on data preprocessing for data stream mining: current status and future directions”, Neurocomputing, Vol: 239, pp: 39-57, May 2017.
31. Syed Khairuzzaman Tanbeer, chowdhury Farhan Ahmed, Byeong-Soo Jeong, Young-Koo Lee,”Sliding windowbased frequent pattern mining over data streams”,Inf. Sci.,Vol:179,No:22,pp:3843-3865,Nov 2009
32. Yunyue Zhe, Dennis Shasha,”StatStream Statistical monitoring of thousands of data streams in real time”,in Proc. VLDB,Hong Kong,pp:358-369,2002.
33. Jaein Kim and Buhyun Hwang,”Real time stream data mining based on CanTree and Gtree”, Inf. Sci.,Vol:367-368,pp:512-528,Nov 2016.
34. Heungmo Ryang and Unil Yun, ”High utility pattern mining over data streams with sliding window technique”, Expert Syst. Appl.,Vol:57,pp:214-231,Sep 2016.
35. Jianneng Cao, Panagiotis Karras, Chedy Raissi and Kian Lee Tan, “Uncertainity inference proof transaction anonymization”, in Proc. VLDB, Singapore, pp: 1033-1044, 2010.
36. Manolis Terrovitis, Nikos Mamoulis and Panos Kalnis, “Privacy preserving anonymization of set valued data”,in Proc. VLDB, Auckland, New Zealand,pp:115-125,2008.
37. Yeye He and Jeffrey F Naughton, ”Anonymization of set valued data via top down, Local Generalization”, in Proc. VLDB Lyon, France,pp:934-945,2009
38. Gabriel Ghinita, Yufei Tao and Panos Kalnis,”On the Anonymization of sparse high dimentional data”,in Proc. ICDE,Cancun,Mexico,pp:715-724,2008

AUTHORS PROFILE

Mr. Jayendra kumar is currently Research Scholar at Computer Science and Engineering Department Koneru Lakshmaiah Education Foundation (Deemed to be University) Vaddeshwaram,Guntur Dist.,AP . He obtained M.Tech CSE from JNTU Hyderabad. His research interest are Data Mining, Internet of Things and Machine Learning . He is Life Member of Computer Society of India.

Dr. Anitha Raju, received a Ph.D. degree in Image Processing and Pattern Recognition from Sri Padmavathi Mahila Visvavidhyalayam, Tirupathi.She is a Woman Scientist sponsor by DST from Govt. of India under the scheme of WOS-A. She is Assoc. Prof. Dept. of Computer Science and Engineering in Koneru Lakshmaiah Education Foundation. Her research interest is the Internet of Things, Image Processing and Machine Learning.