Published by :
http://www.ijert.org

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 10 Issue 01, January-2021

Survey Paper on Various Challenges in Data Stream Mining

Aditya Ghongade [1] ,Adriel Dsa [2] ,Harsh Munot [3] ,Parth Lokhande [4]
*Department of Computer Engineering,
Pimpri- Chinchwad College of Engineering, Pune, India

***Abstract:*- In recent years, advances in hardware technology have facilitated new ways of collecting data in a continuous manner. In many applications such as sensor networks, internet traffic the volume of such data is so large that it may be impossible to store the data locally. Eventually, even when the data can be stored, the volume of the incoming data may be so large that it may be impossible to process any particular record more than once. Therefore, many data mining techniques become more challenging. The progress in hardware technology has advanced and hence made it possible for organizations to store and record large streams of transactional data. Such datasets which continuously and rapidly grow over time are referred to as data streams. Data Stream Mining is the process of extracting knowledge from continuous flow of data which comes to the system in a stream. After a lot of research, data mining has become a well established field now, the data stream problem poses a number of challenges which are not easily solved by traditional data mining methods.This paper proposes various challenges in the data stream and also provides different ways to handle them efficiently.**

Keywords- *Concept drift, Data stream mining, Delayed and missing labels, Skewness*

I.INTRODUCTION

Data streams are high-speed, continuous flow of data.In recent years, advances in this technology have facilitated the ability to collect data continuously. Basic necessities in everyday life such as using a phone or browsing the web,the stock market leads to automated data storage. Similarly, advances in information technology have led to large amounts of data across IP networks. In many cases, these large volumes of data can be mined for interesting and relevant information in a wide variety of applications. When the volume of the underlying data is very large, it leads to a number of challenges:

A. Concept Drift

Concept Drift is a challenge faced while dealing with Data Stream Mining. Concept drift occurs mainly when the data generated is continuously changing for example weather forecasting, online customer buying pattern, etc. These changes are difficult to figure out and are mainly hidden. These changes mainly contribute to what is called concept drift. Concept Drift is a challenge which needs to be handled during Data Stream Mining and affects the previous stored data i.e. makes it useless by shifting the so called concept. There are mainly three types of concept drift namely abrupt concept drift which means sudden changes in values incremental Concept drift which is slow changes in values and last that is gradual concept drift which changes the class distribution of the values.

Concept Drift is an important challenge that one must face while dealing with Data Stream Mining. This means that we should not let the historic data be affected due to the newly generated data. If concept drift is not handled properly it may lead to false analysis or predictions which may affect the efficiency of the program resulting to inaccurate results. Hence to avoid such kind of problem we have to design such an algorithm that not only detects but it also adapts this concept drift regardless its type and furthermore should be capable of predicting when a concept drift will occur which will improve the efficiency as well as the accuracy of the algorithm.

B. Skewness:

Data mining requires gathering and proper selection of data. The skewed data are widely used terminologies that refer to something that is out of order or distorted on one side. We can say that skewness is undistributed data. The primary reason skew is important is that analysis is based on normal distributions. Normal distribution can be termed as the ideal state for our data but in reality we get some distortion in our data set that leads to skewness. So it needs to be handled efficiently and properly.

There are two types of skewness that are positive and negative skewness. In positive skewness data is centered to the right side of median and in negative skewness data is centered to the left side of median.

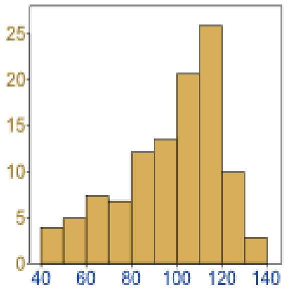

Fig. 1. Negative skew

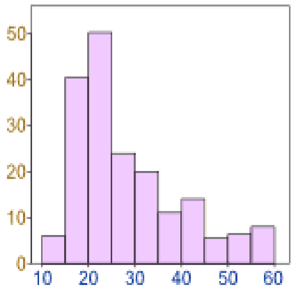

Fig.2. Positive skew

It is important to solve skewness, as data should be uniform and it should be evenly distributed for proper functioning of algorithms as if data is bended towards one region most of the algorithm fails at boundary. So removing skewness from data is one of the challenge that needs to be solved while working on data mining . Skewness tells us a lot about where the data is situated. We get a very good understanding of data by properly studying skewness. In data stream data will generate continuously so if skewness is not handled the point it arrives it will take that error forward and may result in malfunctioning of the algorithm.

C. Delayed and incomplete information:

The amount of data generated and consumed

has increased to a great extent in the past few years. With the internet facilities, this data has even reached remote places. Each second, data is being produced and this data is continuous in nature. This data being generated has to be processed for future use. Many technologies are being used to handle such data. Stream data can include data such as messaging, sensor network, stock market, network traffic,etc. The general technique for stream data mining is to use a pipeline solution in which features of interest are extracted and also online data mining techniques and incremental methods are used for mining. The main disadvantage is that the assumptions made using these techniques are not always correct, they are most often delayed or incomplete. But the techniques used assume that the data is complete and thus affects its performance.

Incompleteness in the data may be due to storage issues and connectivity issues. Delayed information can be attributed as latency where higher latency will cause higher delay.

In the case of latency, the value of the preceding instance's target variable is not available before the subsequent instance has to be predicted. On evolving data streams, this is a mere problem of streaming data integration between feature and target streams. It means that feedback of the current prediction is not available to improve the subsequent predictions, but eventually will become available for much later predictions. Thus, there is no recent sample of labeled data at all that would correspond to the most-recent unlabeled data. And thus it becomes necessary to handle this latency or delayed information.

The need to handle incomplete information is the problem which occurs that the frequency in which missing values occur is unpredictable, but largely affects the quality of imputations, and to select the best imputation technique.For eg: Censored data.

Data Stream mining comprises many such challenges, such as infinite data length, handling of the steady stream of information, effective result presentation from the mining data, scalability of the systems, as well as privacy preservation of the data in the process of extraction from the stream of data. We have studied concept drift, skewness, handling of delayed and incomplete information in our survey. This paper is organized as follows. In Sect.II, we are reviewing the literature topics we studied to handle these challenges. In Sect. III, the conclusions are provided. The main objective of this paper is to understand different challenges and provide efficient solutions to handle them.

II. LITERATURE SURVEY

The Objective of [1] is to provide an efficient algorithm that not only solves the problem of CBM(Condition Based Maintenance) which is the time when a specific machine starts to function abnormally but also works with offline classifiers using ensemble learning and gives an efficient and accurate result.

Industrial Internet of Things (IIoT) devices or sensors are installed in production machines which collect big data on machine conditions and transmit it to the cloud located in the center of the factory. Then the system implements various condition-based maintenance (CBM) methods to predict the time point when machines start to be operated abnormally and to maintain them or replace specific components in advance so as to avoid manufacturing difficulties i.e. to avoid manufacturing defective products. In practice, most companies may not have a sufficient budget to establish a sound infrastructure to support real-time online classifiers, but may have off-the-shelf offline classifiers in their existing systems. This paper proposes an algorithm which is based on ensemble learning technique which also supports online classifiers that cope with three-stage CBM with concept drifts and imbalance data which comprises of three stages which handle specific function of this algorithm. Stages 1 & 3 which are training an ensemble classifier and creating a new ensemble which then employ an improved Dynamic AdaBoost.NC classifier and the SMOTE method to address imbalance data; and Stage 2 which is detecting concept drifts in imbalance data implements an improved LFR (Linear Four Rates) method.

Generally, production managers are based on their domain knowledge and previous experiences to judge the time point when a control component starts to perform abnormally, but the judgement may not always be precise. If the managers judged time is earlier, the cost increases. If the managers judged time is too late , the machine could be malfunctioned. Ensemble learning is a supervised machine learning method. The idea of ensemble learning is to consider a group of experts which determine a final result according to a certain voting scheme of all the experts.

Ensemble learning uses the forecast results from multiple models into a single forecast result. Some works have shown that ensemble learning often performs better than any single forecasting model. At a certain time point by considering a data instance in which each data point has a feature vector X and a class label y in the feature space. The joint distribution of these feature vectors and class labels is called a concept. When an original joint distribution changes to a new joint distribution it is called concept drift.

The authors Minku and Yao proposed an ensemble learning algorithm based on diversity for dealing with drifts also called DDD. When a concept drift is detected, the DDD is trained to train two classifiers for each high and low diversities which helps to adapt to concept drifts. With continuous advances in manufacturing technologies, machine components become increasingly precise. As these machines are so precise the show less malfunctioning of the machine and hence the ratio of fault data and normal data varies a lot. Such an imbalance data problem already exists in the real world some of them are credit card frauds, disease diagnosis, risk management, and fault detection in manufacturing productions.

The DAMSID framework is based on the DDD consisting of three stages: ensemble learning, drift detection, and drift adaptation. In the DAMSID, Stage 1 adopts the Dynamic AdaBoost.NC ensemble learning method which uses the SMOTE method to address the problem of imbalance data; Stage 2 uses the LFR to detect concept drifts; and Stage 3 uses the ensemble learning method to create a new model to adapt to the detected concept drift.

The conclusions derived from this paper are (i) CBM analyzes the machine based on certain conditions to predict the time when the machine starts to function abnormally and to replace or maintain it in advance.(ii) As most of the classifiers can be trained offline hence DAMSID Ensemble Learning based algorithm is used which not only uses offline classifiers but also addresses CBM concept drift and imbalance data (iii) By using this DAMSID algorithm the accuracy can go upto 94% and beyond .

The authors Ravi Kishan Surapaneni, Sailaja Nimmagadda, Roja Rani Govada in [2] have proposed an efficient optimization scheduling algorithm for handling delayed and incomplete information. The proposed algorithm is used to effectively schedule the tasks in data streams. Measures such as volatility, Hurst exponent and distance are used to select task from the data stream. Enthalpy value is then computed based on these features and this enthalpy value is taken as feedback ID. Finally, krill herd optimization algorithm is used for optimum scheduling.

The following block diagram *Fig.3* shows the complete flow of the proposed algorithm:

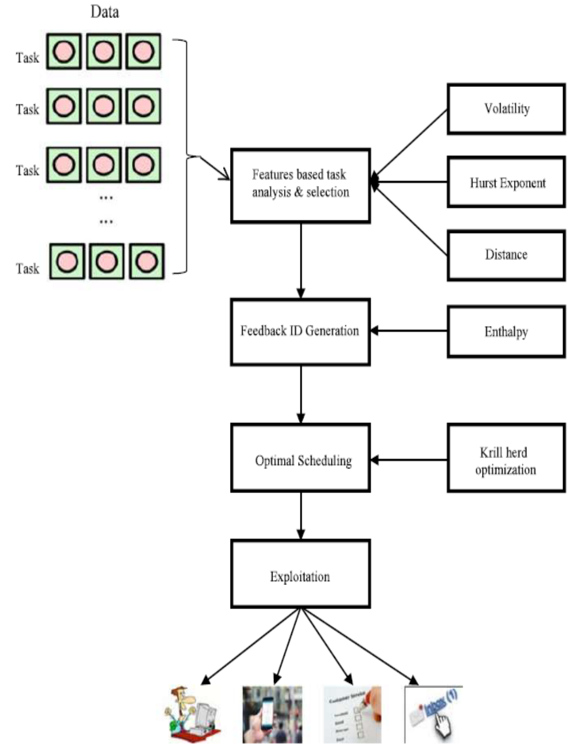

Fig.3. Block diagram of proposed optimal scheduling in big data streams

The data in big streams and number of tasks in streams are analyzed and the tasks are selected for scheduling based on the features such as Volatility,Hurst exponent and Distance.

Volatility Measure: Volatility measures the time differences of data streams.

Hurst Exponent (H): The long time memory nature of time arrangement is evaluated in Hurst Exponent.
H = 0.5 implies time arrangement is uncorrelated.
H > 0.5 implies information with long-extend correlations.
H < 0.5 implies the existence of long-range anticorrelation in data.

Distance: Recognizing the time arrangement of huge information streams is distance measure.

Feedback ID Generation:

It is generated based on Enthalpy Measure.

Here, the enthalpy is identified utilizing the computed measures for task analysis and selection.

E_y(Enthalpy) = V_y(Volatility Measure) + Z

Z = (D * H)
Where, D= Distance, H= Hurst exponent

Krill Herd Optimization Algorithm:
It is an iterative heuristic strategy. The below flow chart *Fig.4* explains the flow of the algorithm.

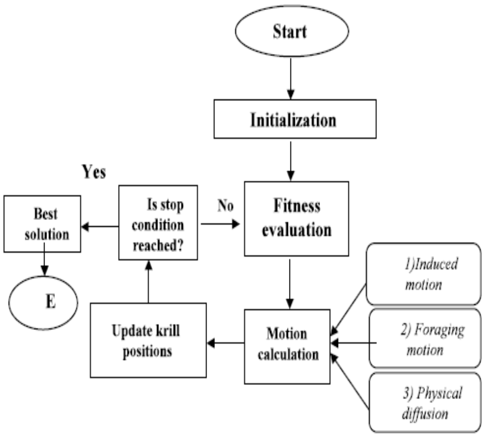

Fig.4. Krill herd flow diagram

Based on Feedback ID generated is initialized. Then Fitness value is calculated from each krill individual and their position. The algorithm loops from best to worst krill individuals.

After that, induced movement, foraging, and random diffusion are evaluated. Then every individual krill value is calculated and is used to update the krill position. If the evaluation is not satisfied then the population is again sorted from best to worst and again the loop continues.

This optimal scheduling algorithm effectively handles delayed and incomplete information. It also outperforms many algorithms on the basis of computational time, schedule time, and throughput.

The authors Tiago Pinho da Silva, Vinicius Mourao Alves Souza, Gustavo Enrique Almeida Prado Alves Batista and Heloisa de Arruda Camargo in [3] have proposed a fuzzy classifier for data streams with infinitely delayed labels called FuzzMiC. This algorithm generates a model for classification of continuous streams based on fuzzy micro-clusters that provides class boundaries to handle infinitely delayed labels.

This paper proposes a classification method to handle the most challenging scenario of extreme latency (delay). To obtain these models, they have used Supervised Fuzzy Micro-Clusters (SFMiC) as summarization structure for the classification for incoming examples. Based on the memberships values, the algorithm associates a class label with a new example. FuzzMiC is an approach for classification where extreme latency and incremental changes occur. The process is separated into two phases like offline and online phases. In the offline phase a decision model is learned from an initial labeled set of examples. Later, in the online phase, new unlabeled examples from the stream are incrementally classified in one of the known classes.To handle noise and flexibility, they proposed a method that uses Fuzzy clustering algorithm to create an initial classification decision model composed of summarization structure. And then through the online phase, the method makes use of these structures to classify new examples.

And thus this work provided a classifier to handle labels under extreme latency and also obtained promising results in some datasets.

The Authors Rojina Deuja, Krishna Bikram Shah in [4] focuses on the missing data in today's trending topic that is social media.Social media is known to generate such kinds of streams. Hence this has become an emerging topic over the past years.And hence huge data challenges are generated to deal with huge amounts of data. Dealing with billions of user's data streams is a hectic task. We bring out social data stream mining methods to process it. We then check the most trendy topic in social media data stream mining to do the depth research of the topic. The data streams come from an array of different sources like social media, business transactions, online searches, phone talks, network trafficking .

The patterns and trends are derived from raw social media data streams which can be termed as Social data media mining. The data required comes from two ways: one is by human analysis and other is by using software programs which are used to check through large amounts of data.. The data to be check is plenty and the time span is very less.There are few ways to analyse them:

i) Memes

Meme is some data that is shared between people using it in any social media through their social networks ie through images. This can be used to get important data and also to understand the admin and his followers on all social networking sites. When memes get trendy it can be useful to many people for businesses,projects,film industry etc.

ii) Maintaining Data Quality

a) Interruption of data present in the social networks is termed as Noise. Signals which are generated in the social stream mining are very high.. The main area of originating the disturbance in social media data are:

1) Fake/ Unused accounts: Plenty of social media accounts are not maintained or recognized by the person of its true identity. Correspondingly, many accounts are created in different social media sites that are created by the admin but never used again.

2) Spam: Spams are mainly divided by harmful links, virus generation attempts,messages. Spams are very different and their looks make user's feelings for their business strategies.

3) Deceptive content: It contains data that is generated with a key to deceive the user and passed on as a unique key of data.

4) Duplicate data: Duplicate data are formed when there is data which is exactly similar to the existing data . Multiple accounts of the same user or Retweets, And Shares And Forwards, can lead to replicate the same amount of data.

b) MissingData: Missing data is that data which has the missing values. The missing information can be that of an admin or the alternative users.eg, People try to avoid disclosing their profile information or their age by trying not to disclose it or by putting the wrong details. Missing values generally can restrict the capacity to describe the large network regarding the actors and the ones next to them.Two forms of missing data i.e. node level and tie level. Node level is the data which was missing of the person or as if it was never created at all. Node level missing data is divided by the actor who does not take part in the operation. The tie level missing data basically occurs when any connecting unit of the social contacts of the current units and then checking the availability of other units. The K-tree model was used to accurately estimate the properties of complete cascading with up to 90% data missing.

Analysing Data Streams

Community Extraction: Community can be divided as a group of people sharing their common ideas in any social group.To mine this data from a certain social service group, it is necessary that the communities of users in that particular social network are recognized by telling us who they are.

Social Stream Mining and their Applications:

GoogleHotTrends: Google Trends is a Google Inc. Software that is free to everyone and can be very useful. It is mainly used to calculate how often a particular search will be carried out from all the searches across the globe.

Cloud4Trends: Cloud4Trends may be a system that identifies trends and relevant topics in social networks through the real-time.

The conclusions derived are that Social media has become an important part of the lives of all human beings all over the globe. The data which is generated from social sites possess great assets for companies, businesses, scientists and individuals. To get the amount of social media streams, new data mining techniques need to be developed.. This will give us the assurance of the huge amount of information from social media users and get benefits from it.

[5] discusses massive data management that needs scalable solutions.Data is increasing enormously day by day and to manage that data is very difficult. PTBSH is an extension of the present technique PTSH(Partition Tuning-based Skew Handling). But this approach has some restrictions and shortcomings: Partitions must be handled in a very continuous and distributed mode only. PTSH cannot distribute the information as it takes very high shuffling time. To beat the shortcoming of the present techniques we've proposed a partitioning algorithm (PTBSH).

For example : (R1,R2,R3,R4, R5) repartitioning are going to be handled in continuous mode only, so results are going to be (R1,R2), (R3,R4) &(R5) or (R1,R2,R3),(R4, R5) it can not be like (R1, R4), (R2,R5), (R3). This system only focuses on continuous partitioning schemes. Continuous partitioning is ensured by virtual repartitioning in spilled files i.e partition of already partitioned files. This scheme cannot handle the asynchronous mode of the attributes. The communication cost of this scheme is additionally very high during the shuffling phase. PTSH cannot handle the choice of pairing of the attributes from the information. Its time complexity is additionally very high. But PTBSH can be added in any manner either continuous or alternative. As well as, it ensures less overhead because the shuffling time is smaller than PTSH. For the accurate classification they used the concept of bootstrap aggregation (ML algorithm for better stability).

This is the algorithm which is employed to style this partitioning technique PTBSH. Where K(key), A(sequence of continuous combination),B(Sequence of the alternate combination), D (d1, d2......di) presents the subsequence of the continual combination of the attributes whereas C(c1, c2.....ci) presents the subsequence of the alternate combination of the attributes. Sum represented by “S”.

A few distributions of data, for instance, the Bell Curve, are symmetric. This means that the proper and therefore the left parts of the distribution of information are immaculate representations of every other.This is an ideal scenario but not every distribution of data is symmetric. It is observed that data skew emerges out of the physical properties of things and hotspots on subsets of the complete domain.

Cov =std Dev/mean *100

Due to the utilization of fairness and accurate classifying techniques name as, bagging for the distribution of the info at reducer side. The coefficient of variance achieved by the PTBSH is better than other compared methods. Also the average runtime of PTBSH is less than all other techniques like HADOOP,LEEN,CLOSER etc.

III. CONCLUSION

Streaming data however has become a daily need of the current world and thus to handle its challenges is obviously a major work. In this paper we have covered a few of the challenges regarding data stream mining and also the need to overcome them.In the literature survey we have explained the papers based on our challenges and the algorithms which are used. The scope of this field is large and will surely have a great scope in near future.

IV. REFERENCES

[1] Lin, C.-C., Deng, D.-J., Kuo, C.-H., & Chen, L. (2019). Concept Drift Detection and Adaption in Big Imbalance Industrial IoT Data Using an Ensemble Learning Method of Offline Classifiers. IEEE Access, 7, 56198–56207.

[2] Surapaneni R.K., Nimmagadda S., Govada R.R. (2020) Handling Incomplete and Delayed Information Using Optimal Scheduling of Big Data Stream. International Conference on Intelligent Computing and Smart Communication 2019, 147-157.

[3] Vera-Rodriguez, Ruben; Fierrez, Julian; Morales, Aythami (2019). A Fuzzy Classifier for Data Streams with Infinitely Delayed Labels. [Lecture Notes in Computer Science] Progress in Pattern Recognition, Image Analysis, Computer Vision, and Applications Volume 11401, 287-295.

[4] Deuja, R., & Shah, K. B. (2019). An Insight on Social Media Stream Mining. SCITECH Nepal, 14(1), 36–43.

[5] Meena, K., & Tayal, D. K. (2018). Partition Tuning based Bagging technique to Skew Handling. 2018 Second International Conference on Inventive Communication and Computational Technologies (ICICCT).