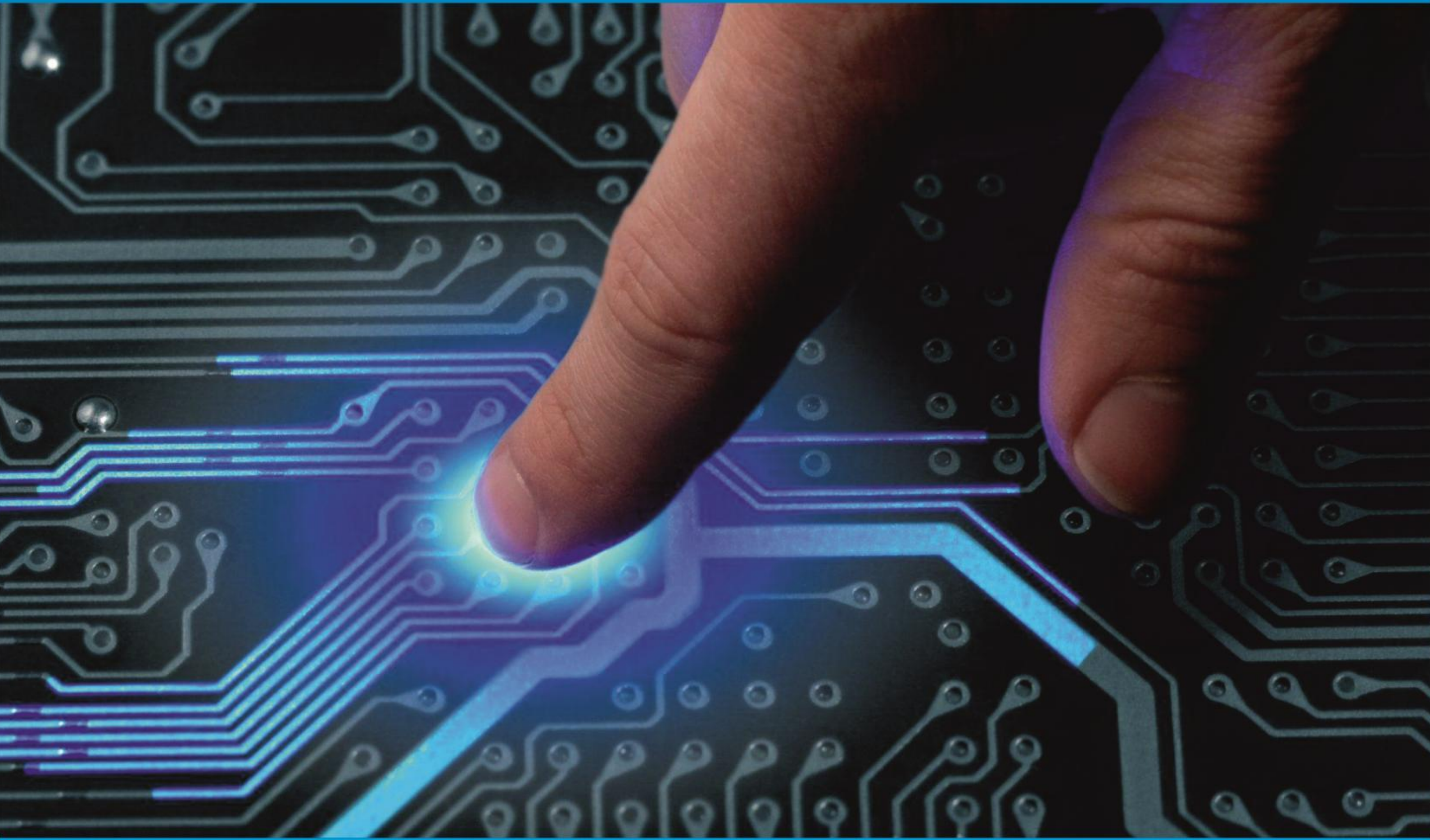
IJIRCCE

e-ISSN: 2320-9801 | p-ISSN: 2320-9798



INTERNATIONAL JOURNAL OF INNOVATIVE RESEARCH

IN COMPUTER & COMMUNICATION ENGINEERING

Volume 9, Issue 4, April 2021

ISSN INTERNATIONAL STANDARD SERIAL NUMBER INDIA

Impact Factor: 7.488

International Journal of Innovative Research in Computer and Communication Engineering

| e-ISSN: 2320-9801, p-ISSN: 2320-9798| www.ijircce.com | |Impact Factor: 7.488 |

|| Volume 9, Issue 4, April 2021 ||

| DOI: 10.15680/IJIRCCE.2021.0904172 |

Data Stream Mining On Real Time Application

Prof. Rahul. A. Patil[1], Chetana Chaudhari[2]

Assistant Professor, Dept. of Computer Engineering, PCCOE, Pune, India[1]

PG Student, Dept. of Computer Engineering, PCCOE, Pune, India[2]

ABSTRACT: Data Stream Mining is the way toward removing data structures from persistent, quick information records. Data streaming is a sequence of unlimited, real-time data objects with a very high data rate. Examples of data streams include computer network traffic, web searches, phone chats, ATM transactions and sensor data. Integration is the process of organizing things into groups with similar members in some way. The naïve Bayes algorithm for dealing with big data by keeping a calculation that is part of the amount of data seen so far in the count-min diagram. A typical study on the conceptual acquisition of online classification with a focus on monitoring student performance rate.

KEYWORDS: Data Stream, Data stream mining, Clustering, Naïve Bayes, concept drifting, Hoeffding tree, Decision tree, Classification.

I. INTRODUCTION

Data Stream Mining is the way toward separating data structures from nonstop, quick information records. Mining data distribution is concerned with the removal of information structures represented in models and patterns in non-stop information streams. Data mining research has found a high level of attractiveness due to the importance of its applications and the increasing popularity of broadcast data. Utilizations of data stream examination can shift from basic logical and cosmic applications to significant business and monetary ones. The biggest hurdles in questioning data streaming are the need for unlimited memory and high data rate. Therefore, the calculation time for each data item should be less than the data rate. In this way, the calculation time per information component ought to be not exactly the information rate. Additionally, it is hard because of unbounded memory prerequisites to have a precise outcome. In many data mining applications, the purpose is to predict the category or number of new conditions in the data stream and to provide specific information about class membership or previous status values in the data stream.

Data stream is a high-speed continuous flow of data from diverse resources. The sources might include remote sensors, scientific processes, stock markets, online transactions, tweets, internet traffic, video surveillance systems etc. Generally these streams come in high-speed with a huge volume of data generated by real-time applications. Data streams have interesting attributes when contrasted and customary datasets. They include potentially infinite, massive, continuous, temporarily ordered and fast changing. Storing such streams and then process is not viable as that needs a lot of storage and processing power. For this reason, they are to be processed in real-time in order to discover knowledge from them instead of storing and processing like traditional data mining. The processing of data streams therefore presents challenges in terms of memory and system performance.

II. RELATED WORK

Fulfillment of data accepts that the genuine upsides, all things considered, that is of highlights and of the objective, are uncovered in the long run to the mining calculation. The issue of missing qualities, which compares to deficiency of highlights, has been talked about widely for the disconnected, static settings. A recent survey is given in [4].Nonetheless, just couple of works address information streams, and specifically advancing information streams.

Paper 1: FarnazAnsarifar and Ali Ahmadi had proposed the “A Novel Algorithm for Adaptive Data Stream Clustering”. In this paper the author proposes the Adaptive data stream Clustering, data stream, clustering, distributed clustering, data stream clustering, data stream mining, Apache Spark techniques. The proposed adaptive STREAM method that determined number of clusters in every chunk. The limit of the STREAM calculation is that it isn't especially touchy to advancement in the fundamental information stream.

Paper 2: MarouaBahri, SilviuManiu and et al had proposed the "A Sketch-Based Naive Bayes Algorithms for Evolving Data Streams". Data stream classification, Naive Bayes, Count-min sketch, hashing trick, Concept drift, Sketch NB algorithm, Adaptive Sketch NB algorithm these algorithms are proposed in this paper. The proposed Sketch NB calculation store information with top notch approximations and the AdaSketchNB calculation stretches out the SketchNB calculation to manage developing information utilizing an idea float system. The Naive Bayes certainly accepts that every one of the traits are commonly free.However, all things considered, it is practically unthinkable that we get a bunch of indicators which are totally free.

Paper 3: Jorge Casillas and et al had proposed the Drift detection in histogram-based Classification, straightforward data stream prediction drift detector algorithm that are used in this paper. The author proposes the histogram-based straightforward Prediction method's that are efficiency in time and memory is extraordinary, which enables it to process thousands of samples per second. The future work is to analyse the method in problems with more than two classes and to study the impact of this drift detector on the learner's performance.

Paper 4: Maciej Jaworski, Piotr Duda and et al had proposed the "New Splitting Criteria for Decision Trees in Stationary Data Streams". Classification, data stream, decision trees, impurity measure, splitting criterion, Online decision tree algorithm these algorithms are used in this paper. Online decision tree algorithm is compared to other algorithms that requires less effort for data preparation during pre-processing and missing values in the data does not affect the process of building decision tree. The future of these paper is to encourage researchers dealing with stream data mining to perform simulations with other values

Paper 5: Zakaria El Mrabet, Daisy Flora Selvaraj and et al had proposed the "Adaptive Hoeffding Tree with Transfer Learning for Streaming Synchrophasor Data Sets". Hoeffding tree, transfer learning, PMU, Oscillations, transfer learning-based hoeffding tree, transfer learning Hoeffding Adaptive Tree Algorithm (THAT) these different algorithms that are used under this paper by author. So the Transfer learning Hoeffding Adaptive Tree Algorithm does not require loading the entire data into memory to build the decision tree model, and thus suitable for real-time processing. The future of this paper is to train the model with recent history holding shorter signatures and to adapt the concept drift in real time.

III. TECHNICAL BACKGROUND

The data stream mining techniques came into existence. They mine frequent patterns in stream data to discover knowledge from huge amount of data for data analysis and decision makingSome information stream mining calculations have pre-processing stage while some different calculations don't have it.

A. *Clustering:*

Clustering can be defined as the identification and classification of objects into different groups, or more precisely, the partitioning of a data set into subsets (clusters) so that the data in each subsetshare some common trait of similar classes of objects (Fig. 1) [2].

1. *Adaptive Stream Clustering Algorithm:*
   Because of huge number of information stream applications, its grouping is a significant method in information mining and information revelation. Stream clustering algorithms is to gain useful knowledge from streams in real-time. STREAM is a data stream clustering algorithm which divides data into chunks, cluster the chunks and, then, again cluster the obtained centres. A significant impediment of the STREAM calculation is that it isn't especially touchy to development in the fundamental information stream.That's why We introduced adaptive STREAM method that determined number of clusters in every chunk. We apply offline clustering to the result of cluster centers obtained from each chunk. This algorithm is as below

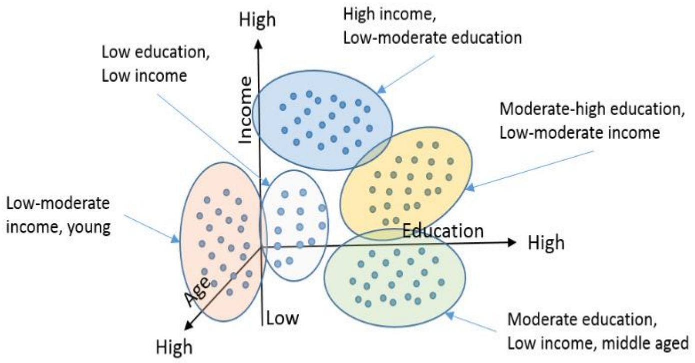


Fig.1 K means clustering[2]

```
Dividing the data stream into chunks d1, d2, ……dr, ….
for each chunk d of data stream D do
Each chunk is d= {a0, a1…., an} with each $a_i$being an array of
length l
d = transpose(d)
pdf [] = 0
tps [] =0
for each row of transposed matrix d
for i← 1 to l do
Cluster containing the PDFs of each element
pdf[i] = kernelDensityEstimation(d[i])
Cluster containing the number of turning points of the PDF
tps[i] = numberOfTurningPointInPDF(pdf[i])
end for
new_k= mean(tps)
totalCenter= kmeans (new_k, d)
end for
return totalCenter
```

B. *Classification:*

Characterization models portray information connections and foresee values for future perceptions. Grouping is the undertaking of learning an objective capacity that guides each quality set X to one of the predefined class names Y. There are distinctive grouping procedures, to be specific Decision Tree based Methods, Rule-based Methods, Memory based thinking, Neural Networks, Naïve Bayes and Bayesian Belief Networks, Support Vector Machines.In characterization test information is utilized to gauge the precision of the grouping rules.On the off chance that the precision is adequate, the standards can be applied to the new information tuples. The classifier-preparing calculation utilizes these pre-arranged guides to decide the arrangement of boundaries needed for legitimate separation.

1. *Naïve Bayes Classifier:*

One of the most often used classifiers is naive Bayes [3]. It uses the assumption that the attributes are all independent of each other and w.r.t. the class label uses Bayes's theorem to compute the posterior probability of a class given the evidence. Using Bayes's Theorem one can compute the probability of each class:

$$P(C|A_1, \ldots\ldots A_a) = \frac{P(A_1, \ldots.A_a|C).P(C)}{P(A_1, \ldots\ldots A_a)}$$

where P(C | A1,….,Aa) is the posterior probability of the target class given the attributes, P(C) is the prior probability of the class, P(A1,…., Aa | C) is the likelihood, and P(A1,…., Aa) is the prior probability of attributes.

2. *Sketch Based Naive Bayes Algorithms:*

To reduce memory, we use sketches algorithms that estimate the marginal and the joint probabilities within a Bayesian network are used. Sketch-based techniques summarize massive data streams using limited space by using multiple hash functions to decrease the probability. The principle focal point of our work to expand the stream gullible Bayes calculation to manage enormous information by keeping the fragmentary tallies of the measure of information. The Sketch Naïve Bayes SketchNB calculation, stores information with excellent approximations in the sketch which permits both quick forecasts and uses a negligible measure of room for preparing.

During the classification process, the sketch table will be used in two steps:

- Learning: refreshing the sketch table for each quality each time another case shows up.
- Prediction: retrieving the counts of a given instance, and use them to compute the naive Bayes probability

3. *Data Summarization Techniques*

This section states the two key components of our solutions for mining data streams.

1) Count-Min Sketch:

Applications can now generate data at rates and volumes which cannot be reasonably stored. The Count- Min Sketch (CMS) [3] which is a generalization of Bloom filters used for counting items of a given type, using approximate counts that are theoretically sound.CMS consists of a two-dimensional array of w .d cells of counters, having a width w of columns, and a depth d of rows(Fig. 3.1).

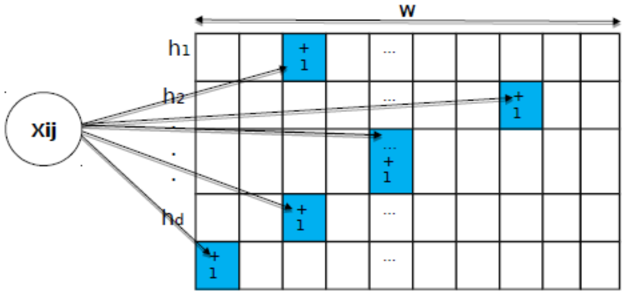


Fig. 3.1 Count-min sketch [3]

2) Hashing Trick:

Hashing Trick (HT), also known as feature hashing [3], is another popular data summarization technique for dimensionality reduction. It is used when dealing with a massive number of features. Hashing trick has been used to make the analysis of sparse and large data tractable in practice. The idea behind the hashing trick is presented in Fig. 3.2, where the sparse feature values are mapped into a lower dimensional feature space.

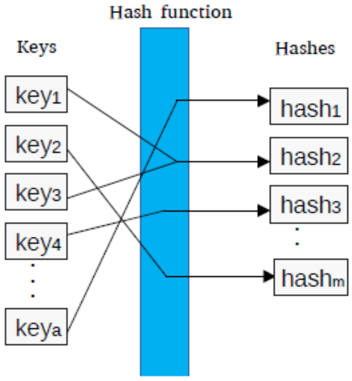


Fig. 3.2 Hashing trick [3]

C. *Transfer Adaptive Hoeffding tree(THAT):*

The Hoeffding Tree (HT) is considered as a standard decision tree used for classification. It uses a Hoeffding [5] bound that relies on the minimum number of arriving samples to build a certain confidence threshold to build trees. Therefore, avoids the entire data to be loaded into memory. The entropy of a given attribute in a training data set is calculated by:

$$Entropy(A) = \sum_{i=1}^{n} - P_i log_2 P_i$$

At that point, the data acquire is registered by:

$$Information\ Gain\ (S, A) = Entropy(A) - \sum\nolimits_{k \in (A)} \frac{|S_k|}{|S|}\ Entropy(S_k)$$

Where k is the value of the attribute A, and S_k is a subset of S where A = K. The other metric considered for evaluation is known as Gini Index, and this index is computed as:

$$Gini(A) = 1 - \sum\nolimits_{i=1}^{n} P_j^2$$

D. *Histogram-based straightforward prediction:*

This float indicator strategy to adequately distinguish varieties in the choice limit that would prompt idea floats. The principle thought is we utilize the clear expectation to screen the elements of the choice limit.We consider the two increments and diminishes of the mistake as idea float in opposition to the regular methodology zeroing in just on decreases of the expectation precision.The point of any float locator is to perceive changes in nonstationary information streams precisely and convenient. To do so, most of research monitors the reduction in the predictive capability as an indicator of changes in the data [1].

E. *Decision tree Algorithm:*

Recently, the demonstration of Hoeffding decision trees are an effective tool for dealing with stream data. For instance, traditional choice trees, for example, ID3 or CART can't be embraced to information stream mining utilizing Hoeffding's imbalance.Along these lines, there is need to foster new calculations, which are both numerically defended and by great execution.In this paper, to overcome this problem by developing a family of new splitting criteria for classification in stationary data streams and investigating their probabilistic properties.

IV. CHALLENGES

Some of the major challenges in data stream mining are handling the continuous flow of data, minimizing the energy consumption, unbounded memory requirement, transferring data mining results, modelling changes of results over time, real time response, and visualization of data mining results [3]. These challenges and some of the methods suggested to overcome them are discussed in the following sections.

1. Handling the continuous flow of data
   This is a data management issue. Conventional data set administration frameworks are not equipped for managing such nonstop high information rate. As an outcome, as a rule it is unfeasible to store all information in tenacious media and in different cases it is excessively costly to arbitrarily get to information on numerous occasions.

2. Minimizing energy consumption
   Enormous quantities of information streams are produced in asset compelled conditions. Sensor networks represent a typical example. Remote sensor organizations (WSNs) have as of late caught the overall consideration because of its gigantic potential for business just as military applications. A WSN consists of low-power, low-cost, and energy-constrained sensors with limited communication and computation ability. These devices have short life batteries.

3. Unbounded memory requirements
   Another feature of data streams is that data are unbounded but storage that can be used to discover or maintain the frequent datasets is limited. AI strategies address the fundamental wellspring of information mining calculations. Most AI strategies expect information to be occupant in memory while executing the investigation calculation.

4. Transferring data mining results
   Knowledge structure representation is another essential research problem. Given that transfer of all data to a central site is not feasible and scalable, nowadays extraction of interesting patterns are done locally and after extracting models and patterns locally from information stream generators, it is vital for move these designs to the client. This reduces the traffic and the load on the central site.

5. Modelling changes of results over time
   In some cases, the user is not interested in mining data stream results, but to know how these results change over time. For example the number of clusters generated by a data stream changes over time, which might represent some changes in the dynamics of the arriving stream.

6. Real-time response
   Since information stream applications are normally time-basic, there are prerequisites on reaction time.For some restricted scenarios like emergency response in flight navigation at air traffic control facilities, algorithms that are slower than the data arriving rate are useless. Present day Air-Traffic Control frameworks give initially a more clear, more complete image of the blockage status of a given airspace. The mouse floating over each track image permits the regulator to see a plenty of information and issue orders that can get transferred to the airplane by means of satellite.

7. Visualization of data mining results
   Perception of customary information mining results on a work area is as yet an exploration issue.Perception in little screens of a Personal Digital Assistant (PDA) for instance is a genuine test.Imagine a businessman at his job analysing the data streaming in from his head office and various point-of-sales on his Personal Digital Assistant. The results of his analysis should be efficiently visualized in a way that enables him to take a quick decision.

8. Fluctuating data rates
   Scientists have to deal with the fact that the data rate of the stream isn't constant, leading to a condition called bustiness, and the patterns of the data stream and scheduling resources are continuously evolving. Most of the data streams available for mining today exhibit changes in underlying process that generate the data.

9. Temporal locality
   By and large, there is an inalienable transient segment to the stream mining measure. This is on the grounds that the information may advance over the long haul. This conduct of information streams is alluded to as fleeting region. Numerous applications, for example, news bunch sifting, text creeping, and report association require constant grouping and division of text information records. The unmitigated information stream bunching issue likewise has various applications to the issues of client division and constant pattern investigation.

V. Conclusion

In this report the theoretical aspects of stream data mining classification algorithms. The proposed adaptive stream method that determined number of clusters in every chunk. A transfer learning technique using Hoeffding tree is proposed for synchrophasor data. The proposed model, called THAT, does not require loading the entire data into memory to build the decision tree model, and thus suitable for real time processing. The naive Bayes classifier using the count min sketch to reduce the memory needed. It is reasonable for true issues as it manages numeric traits and missing qualities. The calculation can be utilized for building more modest or bigger, more precise choice trees and the calculation is very time proficient.

References

1. Jorge Casillas, Shuo Wang, Xin Yao [2018]. Concept Drift Detection in Histogram-Based Straightforward Data Stream Prediction. IEEE International Conference on Data Mining Workshops (ICDMW)
2. FarnazAnsarifar, Ali Ahmadi (2018). A Novel Algorithm for Adaptive Data Stream Clustering, 26th Iranian Conference on Electrical Engineering (ICEE2018)
3. MarouaBahri, SilviuManiu, Albert Bifet [2018]. A Sketch-Based Naive Bayes Algorithms for Evolving Data Streams. IEEE International Conference on Big Data (Big Data)
4. Maciej Jaworski, Piotr Duda, and LeszekRutkowski, Fellow, IEEE2019 New Splitting Criteria for Decision Trees in Stationary Data Streams
5. Zakaria El Mrabet, Daisy Flora Selvaraj, Prakash Ranganathan. Adaptive Hoeffding Tree with Transfer Learning for Streaming Synchrophasor Data Sets. 2019 IEEE International Conference on Big Data (Big Data)

INNO SPACE
SJIF Scientific Journal Impact Factor

Impact Factor:
7.488

INTERNATIONAL
STANDARD
SERIAL
NUMBER
INDIA

निस्केयर
NISCAIR

INTERNATIONAL JOURNAL OF INNOVATIVE RESEARCH

IN COMPUTER & COMMUNICATION ENGINEERING

9940 572 462 6381 907 438 ijircce@gmail.com

Scan to save the contact details

www.ijircce.com