Data Stream Mining Based on the Misclassification Error

Archana P[1], Jayalakshmi[2]
[1]*USN: 1BM14IS400, Department of ISE, BMS College of Engineering, Bangalore, India*
[2]*USN: 1BM14IS407, Department of ISE, BMS College of Engineering, Bangalore, India*

ABSTRACT

In recent years, the amount of data that needs to be analyzed is growing very fast. Potentially unlimited number of data is the cause of creation of a new field of research called data stream mining. By analyzing stream of data elements, one has to face new difficulties, therefore standard approach to the problem of data mining cannot be applied. A new method for constructing decision trees for stream data is proposed. First a new splitting criterion based on the misclassification error is derived. A theorem is proven showing that the best attribute computed in considered node according to the available data sample is the same, with some high probability, as the attribute derived from the whole infinite data stream. Next this result is combined with the splitting criterion based on the Gini index. It is shown that such combination provides the highest accuracy among all studied algorithms.

Keywords—*Data mining, Misclassification Error, Gini index.*

1. INTRODUCTION

As the technology is growing at a faster rate amount of data that needs to be analyzed is growing very fast. To support this a new technology called data stream mining is emerging. Information Stream Mining is one of the zone picking up part of handy criticalness and is advancing at an energetic pace with new techniques, approaches and discoveries in different applications identified with solution, software engineering, bioinformatics and securities exchange expectation, climate estimate, content, sound and video preparing to give some examples. one of the worry is the information used in information mining. With the immense information obtained by the online creates a few sensors, Internet Relay Chats, Twitter sites, Face book, Banking online or Transactions of ATM, the idea of powerfully changing information is turning into a key test, what we call as information streams.

Data Stream Mining is the path toward removing taking in structures from unending, fast records of data. A streaming of data is a asked for progression of illustrations which is in various employments of data stream mining which can scrutinized just not only once ,can be many more few times using obliged figuring and limit capacities. Instances of data streams join PC organize action, phone discourses, ATM trades, web journeys, and data sensors. Data stream mining is a subfield of data mining, machine learning, and learning revelation. For various applications of data stream mining, goal is to suspect the class or estimation of new events in the data stream given some finding out about the class cooperation or estimations of past cases in the streaming of data.

By analyzing stream of data elements, one has to face new difficulties, therefore standard approach to the problem of data mining cannot be applied. One of the difficulties is the need of on-the-spot data analysis. This results from the fact that all data cannot be stored due to the limitation of memory. The second problem is the rate of incoming data. The algorithm used for data stream analysis should be fast to keep up with the incoming data.

In this paper a new method for constructing decision trees for stream data is proposed. First a new splitting criterion based on the misclassification error is derived. A theorem is proven showing that the best attribute computed in considered node according to the available data sample is the same, with some high probability, as the attribute derived from the whole infinite data stream. Next this result is combined with the splitting criterion based on the Gini index. It is shown that such combination provides the highest accuracy among all studied algorithms.

2. RELATED WORK

In the ACM KDD International gathering held in 2010, the creators talk about the issue of finding the top-k visit things in an information stream with adaptable sliding dowagers [3]. The thought is to mine lone the top-k visit things as opposed to revealing all the regular things. However, the vital variable or impediment that develops here is the measure of memory that is required still to mine w.r.t to finding of top-k visit things is as yet a

bouncing element. The creators at long last talks about that there exists however a memory proficient calculations by making a few suspicions.

The creators concentrate on building up a system for ordering powerfully developing information streams by considering the preparation and test streams for dynamic arrangement of datasets [2]. The goal is to build up an arrangement framework in which a preparation framework can adjust to snappy changes of the fundamental information stream. The measure of memory accessible for mining stream information utilizing one pass calculations is less and henceforth there is chance for information misfortune. Additionally it is impractical to mine the information online as and when it shows up as a result of crisscross in speed and a few other huge elements.

The creators examine the strategy for finding most successive things by utilizing a hash based approach [4]. The thought is to utilize say "h" hash capacities and fabricate the hash table by utilizing direct congruencies. Information streams can be grouped into two sorts as: Disconnected information streams and Online information streams.

The strategy for particular esteemed disintegration is utilized to discover the connection between various streams [6]. The idea of SVD was especially used to discover disconnected information streams. Bunching content information streams is one of the points which have developed as essential test for information mining specialists. The issue of spam discovery, email sifting, bunching client practices, subject location and recognizable proof, report grouping are a couple of run of the mill enthusiasm to information mining analysts.

Liu et.al examine on grouping content information streams [7]. The thought is to expand the current semantic smoothing model which functions admirably with static information streams for grouping dynamic information streams. The creators propose two web based bunching calculations OCTS and OCTSM for grouping enormous content information streams. A colossal measure of information is created from web each moment in different structures, for example, interpersonal organizations, information from sensors, confront book and twitter. The rising information from web likewise called as Text message stream which is created from different text applications and web transfer visit. This has turned into a prime point which has turned into an intriguing issue important to the analysts working in the region of information mining and has a great deal of degree to work to be contributed by the examination group.

Shen, Yang et.al, proposed the technique for recognizing the strings in powerful information streams [8]. The paper examines three varieties of single pass grouping calculation taken after by a novel bunching calculations in light of phonetic elements. A strategy for diminishing the dimensionality of gushing information utilizing an adaptable administered calculation is proposed in [9].

The confinements of PCA, LDA and MMC methodologies are examined. The creators call attention to the unsatisfactory quality of MMC for gushing information. An administered incremental dimensionality diminishment calculation is proposed to meet the necessities of gushing informational index. The authors demonstrate that the most referred to come about Hoeffdings bound is invalid [10]. Aggarwal et. al. proposed an alternative tress for information stream mining. Information streams are universal and have over the most recent two decades turn into an imperative research subject. For their prescient nonparametric examination, Hoeffding-based trees are frequently a technique for decision, offering a plausibility of at whatever time forecasts. Be that as it may, one of their primary issues is the postponement in learning progress because of the presence of similarly discriminative qualities. Alternatives are a characteristic approach to manage this issue. Alternative trees expand upon consistent trees by including part choices in the inside hubs. All things considered they are known to enhance exactness, dependability and diminish uncertainty.

3. PROPOSED WORK

In the Proposed method input is given in the form of dataset, once input is loaded, preprocessing of data takes place by removing null attributes and unwanted records, then information is obtained and based on the information gain gini index value is calculated. Output is generated in the form of hybrid tree and tree is generated only for the highest value obtained by the gini index.

1. Dataset Collection

Collect the dataset according to the identified application.

2. Pre-preparing information

First pick a set n dataset which must be gathering together to detect natural structure of the report space utilizing connection. At that point we evacuate invalid words, for example, and, or, this, is and so on we expel the words which has no goal importance from n dataset to lessen the multifaceted nature amid correlation and composing the substance in the wake of expelling words into another record under the organizer stop words.

3. Discovering imbalance or inequality

Imbalance, otherwise called limited contrasts disparity or Hoeff ding-Azuma imbalance. Let X1… Xn be free irregular factors, with Xi taking qualities in some set.

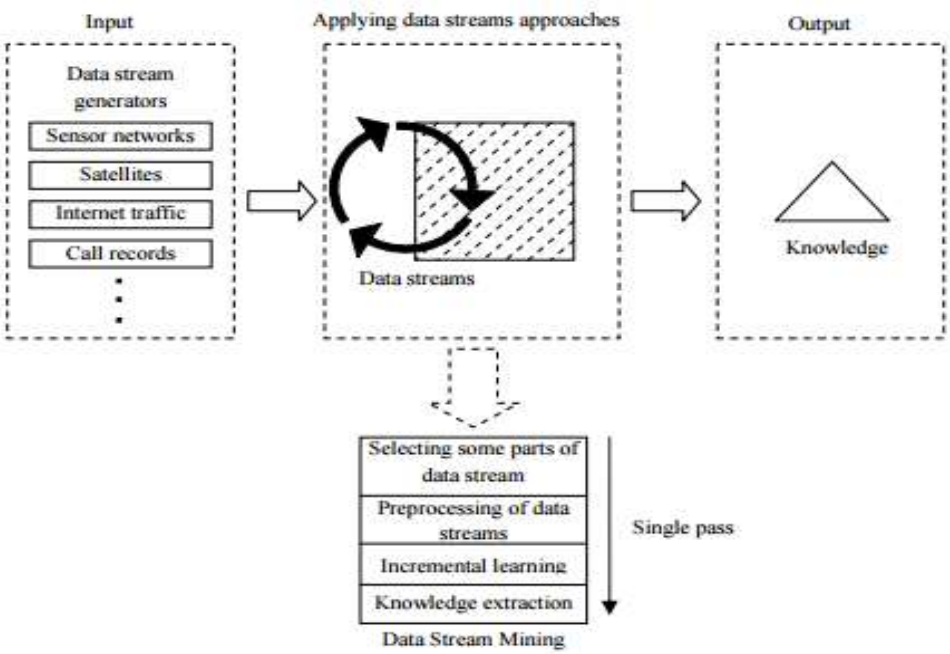


Fig -1: Proposed Architecture

4. Discovering gini record or index

Choice tree learning framework, connected to information streams, has the property that its yield is almost indistinguishable to that delivered by a regular learner.

Data pick up proportion inclinations the choice tree against considering characteristics with countless qualities. So it comprehends the disadvantage of data pick up—to be specific, data increase connected to qualities that can interpretation of a substantial number of unmistakable qualities may take in the preparation set.

This characteristic has a high data pick up, in light of the fact that it particularly recognizes every client, except we would prefer not to incorporate it in the choice tree. Information gain ratio biases the decision tree against considering attributes with a large number of distinct values. So it solves the drawback of information gain—namely, information gain applied to attributes that can take on a large number of distinct values might learn the training set.

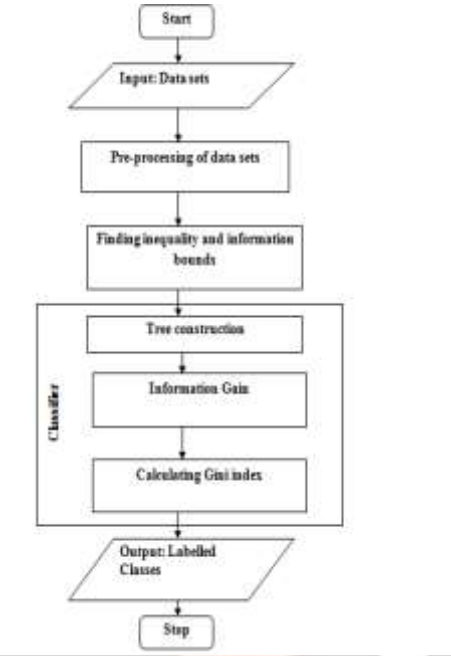


Fig-2: Flow Diagram

5. Decision tree utilizing hybrid algorithm

The quantity of information components N, the CART calculation is a great deal additional tedious than the Hybrid Tree calculation. The preparing time for the CART calculation is a power capacity of N. For the Hybrid Tree calculation the connection is practically straight.

The measure of memory does not rely on upon the extent of a preparation informational index. For the CART calculation the expended memory depends directly on N. The exactness increments with the development of the size N of the preparation informational collection.

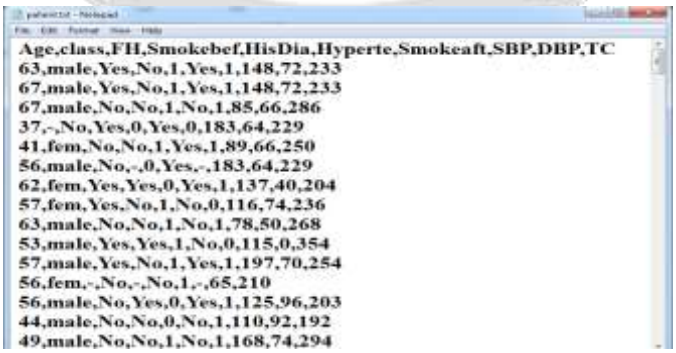
4. EXPERIMENT RESULT

We implemented Proposed work with following assumptions and dependencies:

The following have been used in the development of our project:

- All data sets have the same availability probability.
- Sql server 2000 is used as a back end tool for maintaining the information about data sets, attributes and preprocessed data.
- The framework application requires proper Java environment installation.

As a proof of concept we considered patient dataset from KDD(Knowledge Discovery Dataset). Fig 3 shows the sample dataset considered. Decision tree generated is shown in fig 4.



```
Age,class,FH,Smokebef,HisDia,Hyperte,Smokeaft,SBP,DBP,TC
63,male,Yes,No,1,Yes,1,148,72,233
67,male,Yes,No,1,Yes,1,148,72,233
67,male,No,No,1,No,1,85,66,286
37,-,No,Yes,0,Yes,0,183,64,229
41,fem,No,No,1,Yes,1,89,66,250
56,male,No,-,0,Yes,-,183,64,229
62,fem,Yes,Yes,0,Yes,1,137,40,204
57,fem,Yes,No,1,No,0,116,74,236
63,male,No,No,1,No,1,78,50,268
53,male,Yes,Yes,1,No,0,115,0,354
57,male,Yes,No,1,Yes,1,197,70,254
56,fem,-,No,-,No,1,-,65,210
56,male,No,Yes,0,Yes,1,125,96,203
44,male,No,No,0,No,1,110,92,192
49,male,No,No,1,No,1,168,74,294
```

Fig-3: Patient dataset from KDD

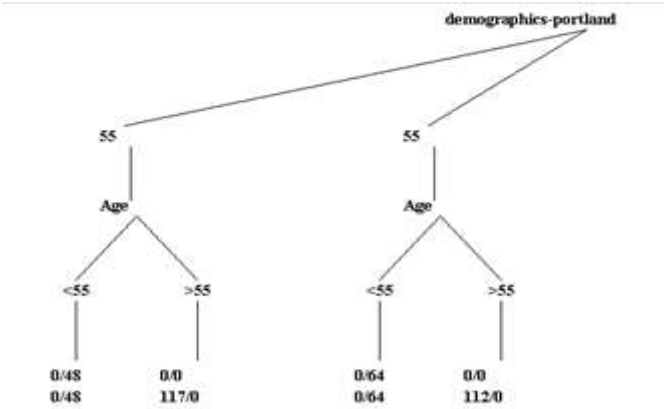


Fig-4: Decision tree generated

5. CONCLUSION

Because of boundless, huge, high volume information getting produced from different applications as information streams it is very normal to deal with them in view of their dynamic, unpredictable and variation nature. The issue of dealing with streams for bunching, characterization and theme recognition is as yet a test and has a wide shot of investigation for information mining analysts to convey their work. In this venture, information stream mining utilizing the half breed calculation characterized which utilizes choice tree. The information stream mining approach has been appeared to create a model that advances incrementally after some time and all things considered is more steady than other arrangement tree approaches.

6. ACKNOWLEDGEMENT

We thank Prof. Abhijith H V, Assistant Professor, Dept. of ISE, BMSCE for his constant guidance, support and encouragement to complete this project.

7. REFERENCES

[1] Bifet A, Holmes G, et.al. Mining Frequent Closed charts on developing information streams. Procedures of seventeenth ACM SIGKDD International Conference on information revelation and information mining, 2011: 591-98.

[2] Charu C. Aggarwal, Jiawei Han, Philip S. Yu. On Demand Classification of Data Streams, in the procedures of ACM KDD'04, USA, 2004.

[3] Hoang Thanh Lam, Toon Calders, Mining Top-K Frequent Items in a Data Stream with Flexible Sliding Windows. Procedures of in the procedures of ACM KDD'10, USA, 2010.

[4] Cheqing Jin et.al. Progressively Maintaining Frequent Items over a Data Stream, in the procedures of CIKM USA, 2003.

[5] Nan Jiang and Le Grunewald. Look into Issues in Data Stream Association Rule Mining, SIGMOD Record, 2006: 35(1).

[6] Sudipta Guha, D.Gunopulos, N. Kaudas. Associating synchronous and offbeat information streams, in the procedures of SIGKDD 2003 held from august 24th - 27th, USA, 2003.

[7] Yu.Bao.Liu et.al. Bunching Text information streams, Journal of software engineering and innovation, volume 23, issue 1, pages 112-128, 2008. [8] Dou Shen,Qiang Yang, Jian-Tuo-Sun, Zheng Chen. String Detection in Dynamic Text Message Streams, in the procedures of SIGIR USA, 2003.

[8] Jun Yan et.al. A versatile regulated calculation for dimensionality decrease on gushing information. Data Sciences, An International Journal, Published by Elsevier, 2006; 176: 2042-65.

[9] L.Rutkowski et.al. Choice trees for mining information streams in view of the McDiarmid's bound. IEEE Transactions on Knowledge and Data Engineering, 2013; 25(6).