International Journal of Engineering and Advanced Technology (IJEAT)
ISSN: 2249-8958 (Online), Volume-8 Issue-4, April 2019

The Impact of Streaming Data Noise Reduction by using Chunk Based Ensemble

Vishal Jaulkar, Nalini N

Abstract*: *The Propelled analytics of data streams is rapidly turning into an important territory of processing the complicated data as the quantity of requesting such applications grows continuously. Web based mining when the input data advancing over period is getting to be most important center problem. In todays real world the data is not stationary but non stationary. The properties of attributed are changing over time. As a result the data stream has concept drift and noise which affects the performance. The aim of the paper is to first present an overview of the challenges in the data streams, followed by the measures to improve the factors that affect the performance.

Index Terms*: *Analytics, concept drift, stream, mining.

I. INTRODUCTION

Information is gushing from all parts of our lives in uncommon sums; at no other time in the historical backdrop of humankind has there been so much data being gathered, contemplated and utilized day by day. On the off chance that there ever was an unrest in business, the substantial measure of data spilling in from our telephones, PCs, stopping meters, transports, prepares, and planes is really it. Not exclusively are organizations gathering a lot of information, however they are likewise utilizing this information to improve clients' encounters and their business choices and procedures. On the off chance that unconvinced about the intensity of this changing and extending mass information, one needs just to scrutinize the changing limit of PCs and the speed at which such changes happen, there is an update at regular intervals. In spite of the fact that utilized in the past to break down different parts of general wellbeing and medication, the data picked up from the information streams is making life on planet earth much better if not seemingly less complex.

Recently the Big data has been gaining much attention due the ability to take key decisions. Big data presents huge number of data sets. But here even though the size of dataset is huge, using big data analysis can be done with much ease and decision making can be improved. Big data along with the applications of machine learning has large number of uses in almost any field. This mix of this enormous datasets collections is very intricate.

Manuscript published on 30 April 2019.
***** Correspondence Author (s)
Nalini N, Asst. Prof. Senior ,Scope Vellore Institute of Technology, University, Vellore (Tamil Nadu) India
Vishal Jaulkar, Scope Vellore Institute of Technology University, Vellore (Tamil Nadu) India.

There are a few difficulties one can look amid this coordination, for example, investigation, information curation, catch, sharing, seek, representation, data security and capacity. One troublesome feature of big data information is the utilization of a wide scope of creative information the executives instruments and systems whose plans are devoted to supporting operational and explanatory mining. As of late, processing information streams with concept change for significant bits of knowledge has turned into a critical and testing undertaking for a wide scope of utilizations including charge card extortion assurance, target showcasing, organize interruption Identification, and so forth. Traditional data mining processing is affected by two main problems which are the continuously increasing size of the data and the data streams impacted by the drift. Web based learning methodologies process each training precedent once "on entry," without the requirement for capacity or reprocessing, and keep up a present theory that mirrors all the training models so far [1]. Along these lines, the learning methodologies take as information a solitary preparing precedent just as a theory and yield a refreshed speculation [2]. We think about web based learning as a specific instance of steady learning. The last term alludes to learning machines that are additionally used to display constant procedures, yet are permitted to process approaching information in lumps, rather than preparing each preparation model independently [3].

Ensembles of classifiers are being effectively used to improve the precision of single classifiers in on the web and steady learning. Notwithstanding, online conditions are frequently nonstationary and the factors to be anticipated by the learning machine may change with time (idea float). For instance, in a data separating framework, the clients may change their subjects of enthusiasm with time. In this way, learning machines used to display these situations ought to most likely adjust rapidly and precisely to conceivable changes. A few methodologies have been proposed to deal with idea float in the course of the most recent couple of years. For example, we can refer to outfit approaches, which make another classifier to each new lump of information and weight classifiers as per their exactness on ongoing information perhaps utilizing a boosting-like component for the learning. Another model is Gao et al's. work [10], which proposes the utilization of unweighted groups, as new information may have a place with an idea not quite the same as the latest preparing information. Road and Kim [11] likewise report that no predictable enhancement for the precision was gotten when utilizing gathering part loads in their methodology.

Regardless of the way that there is a tremendous writing on time cross sectional data processing, most of the current methodologies does not consider whether the sequence of data points taken at timely order is an exceptional sort of streaming [1]. The stream processing is the lot of information perceptions that happen to come consecutively thing by thing [2]. The data stream is the continuously generated data by various kinds of sources. To process such kind of data, it must be analysed in an incremental manner as entire data is not available at once. Data stream has the property of being dynamic in nature. The data streaming is useful in extracting real time patterns from the big data.

But the data stream has concept drift presence.[3]. The greater part of the methodologies intended to time cross sectional data processing investigation are uninformed of Concept Drift. These techniques depend on the primary presumption that time cross sectional data processing ideas are stationary so that the perceptions pursue a fixed and changeless Probability Distribution Model . This supposition, be that as it may, may not hold for a few modern time cross sectional data processing requisites. For instance, the time series preprocessing of the offers of an item may change its conduct because of changes in government guidelines or publicizing efforts. The time arrangement of stock costs of an organization may change its conduct because of changes in political and conservative components or because of changes in the speculators brain science or desires. The class imbalance problem also adds to the issues that affect the performance of the predictive model dealing with the data which is non stationary. Basically the class imbalance is the situation where number of the labelled training data which are positive are less in number compared to the negatively labelled data. Generally, any predictive machine learning model works better if the set of both data points are comparable to each other. Because in the case where the number of positively labelled data is less than the negatively labelled data, most of the prediction will go into the way of negative. For instance. If we are trying to determine whether the transaction of credit card is fraud or not and number of data points that say the transaction is not fraudulent are 1000 compared to the ones that say transaction is fraud are 10. So, mostly the prediction will go towards not fraudulent. This should not be the case and will be determining step for performance of the data stream processing. The two most common ways to solve the problem of Class Imbalance are Under sampling and Oversampling.

II. LITERATURE SURVEY

Following the survey, a few specialists have explored how to deal with concept drift in Supervised Machine learning algorithm issues. Anyway in time series investigation, only a couple of scientists have endeavored to tackle this issue.

The data points measured in timely manner consist of a series and they can be classified by the supervised learning. The supervised learning is the association between the input and the output so that based on the association the output value of the next input can be predicted. As a whole, the output at time Y was predicted based on the input at time X with Y > X. This is the sliding window that is use of earlier timely values to get the next timely values.

The Guajardo et al. has made use of this concept along with the Regression Algorithm. The sliding window trains the input dataset. Whenever it is moves it trains a particular set of data and with the help of regression algorithm, the SVRs try to fir into the data line by minimizing the cost function. As the dataset is trained timely so even if the data is continuously changing, it is taken into the account for building the model for predicting the final output.

Heng and Zubin has proposed a new mechanism for the detection of concept drift. The advantage of the mechanism is that it works on both type of data , data that needs to be feeded all at once as well as streaming data. They have given a framework to detect the drift if present in the data input by determining the data labels belonging to new concept. And as a result the model can be adopted to new changes with improved accuracy and efficiency. The framework proposed by Heng does not depend on the data distribution. So the it works still better even if the static properties of the data set are changing.

In [7], the paper discusses the concept of Hoeffding Tree. The Hoeffding tree name comes from from the concept of Hoeffding Bound. It gives the number of training sets of data sets needed to decide the skill of the classification model. The hoeffding tree using the hoeffding tree bound gives the uncertainty of the estimate. So based on the skill level of the best node in the tree, the machine learning classification can be easily done.

The hoeffding tree algorithm has been extended to Very Fast Decision Tree concept. The main purpose of VFDT is that it does not need complete data set to predict based on the model built. It performs on data set segments. To accommodate the effect of the missing data or incase of streaming data the input continuously changes , so a cache part is proposed. The VFDT also improves the performance of the dataset by predicting the missing values and noise replacement. The time required for building the model for classification is much less in VFDT compared to the others as VFDT uses part of the data set to train and test.

There are many methods proposed for the detecting the concept drift if it is present in the streaming data. The presence of concept drift reduces the accuracy of the model as the properties of dataset attributes continue to change over the time. So the present classification or prediction will become more wrong in future as the time goes by. Also whenever concept drift is detected, the current working model is not accurate and it must be replaced. As a result, it is must to detect the presence of concept drift as in the Real World applications, mostly involve Non Stationary datasets and if not for concept drift could affect the performance of decision making by a huge margin. So to avoid the loss due to concept drift, the prediction model must updated with the dataset distribution that is changing over the time quickly.

The real world applications where the concept drift plays the major part include Cyber Security, Telecommunications and finance.

In [8], a new mechanism has been proposed for the detection of the concept drift. It uses the Ensemble Classifier and Random Decision Tree.

Published By:
Blue Eyes Intelligence Engineering
& Sciences Publication (BEIESP)

The model builds based on the blocks of the data in stream also known as Chunks. The main application of this stated model is that it also considers the various type of concept drifts affected by the data streams with the noise. And even if the amount of the data is very less or incase it is too big, the technique presented by Li et al performs with much accuracy where the other techniques presented would perform with less accurate results.

Learn++.NSE mechanism works in an incremental way. It combines the result of the blocks of the data as they arrive. Like others, it doesn't consider the entire data set at once. It starts with one chunk of the data. When the next chunk become available voting is done based on the accuracy of that chunk whether to include it in the first chunk or not. It uses chunk based ensemble classifier. When the data distribution is not stationary, the Learn ++ NSE is very useful. But in this mechanism, the error rate is of the entire environment. So The Learn++ NSE has been further improved to WMA Learn++ NSE to include sliding window. It counts the average error count of the chunk and not the entire environment. It improves the overall performance of the model built for the Prediction. The WMA Learn++ has many applications in the real world applications compared to Learn++ NSE as it counts vote weight based on the sliding window.

III. CHALLENGES IN DATA STREAM

The data handled by in today's real world applications is in enormous volume. The data is not just stationary but non stationary data also is playing vital role. The stationary data is the type of data with properties that are constant and not changing with respect to the time. In the case of non-stationary data the properties such as mean or variance etc. are constantly changing with respect to time. The non-stationary data is very complex as it is increasing with time. So processing of such large volumes of streaming data could be out of bounds. Some algorithms might not support the continuous input and could be affected. For processing the streaming data, the classification model to be built based on the train datasets must be adopting the changes incrementally so that data results are accurate. Algorithms must be able to process the data in real time. There are many challenges faced while dealing with the non-stationary data. The main problems among them being concept drift, Class Imbalance. The overview of most common challenges that occur in the streaming data is discussed below.

A. CONCEPT DRIFT

One of the most common issue to handle in case of the streaming data is the presence of the concept drift. Concept drift is the shift in between static relation between the input and the output. If y = f(x) stating that based on the value of x, value of y can be determined. In stationary data, this mapping will not change with respect to time. But when the data is streaming, this association is not static. It will change over the time. This change in the association between the input and the output attribute is the Concept Drift. The association could be of hidden type and an unknown one. As the data is continuously changing over time, patterns recognition will be a challenge as it keeps on evolving.

The test issue for stream mining is the capacity to for all time keep up a precise choice model. Besides, they ought to overlook more established data when information is obsolete. In this unique situation, the supposition that models are created aimlessly as per a stationary likelihood appropriation does not hold, in any event in complex frameworks and for huge timeframes.

B. DATA STREAM MINING

Information Stream mining alludes to learning extraction as models or guidelines from advancing information streams. Information Streams have distinctive difficulties like stockpiling for putting away outcomes, questioning, preprocessing and investigating. The principal look into test is to configuration quick and dependable digging models for developing information streams. That is, calculations that just require one disregard the information and work with constrained capacity. The other one is made by the exceedingly unique nature of information streams, whereby the stream mining calculations need to distinguish evolving ideas.

In next area, we will talk about difficulties with respect to the characterization of information stream. The primary worry with information stream grouping is building the classification model and staying up with the latest by every now and again refreshing the model with the latest named information.

C. CLASS IMBALANCE

The class imbalance is the situation where number of the labelled training data which are positive are less in number compared to the negatively labelled data. Generally, any predictive machine learning model works better if the set of both data points are comparable to each other. Because in the case where the number of positively labelled data is less than the negatively labelled data, most of the prediction will go into the way of negative. The class imbalance problem also adds to the issues that affect the performance of the predictive model dealing with the data which is non stationary.

For instance. If we are trying to determine whether the transaction of credit card is fraud or not and number of data points that say the transaction is not fraudulent are 1000 compared to the ones that say transaction is fraud are 10. So, mostly the prediction will go towards not fraudulent. This should not be the case and will be determining step for performance of the data stream processing. The two most common ways to solve the problem of Class Imbalance are Under sampling and Oversampling.

IV. WHY CHUNK BASED ENSEMBLE

When the data is on-stationary, concept drift may be present in the streaming data. The association between data chunks in such concept drift is close. And as a result, the chunk ensemble can be used to test the presence of the concept drift in the data input. In the ensemble learning, multiple classifiers

are trained on different data chunks. Then final outcome is predicted after combination. The chunk based ensemble can be applied on the first data chunk. When the new set of data arrives, the chunk can be updated with new data.

This will make the model to be incremental towards streaming data and robust. It will be efficient as the dynamic relation is getting computed.

Published By:
Blue Eyes Intelligence Engineering
& Sciences Publication (BEIESP)

The chunk based ensemble can be used to detect the novel class present in the data chunks of a stream by training the unlabeled data.

Elegant techniques which combine more than one models have been more well-known than their single model partners due to their less complex usage and higher proficiency [11]. A large portion of these elegant procedures utilize a block based methodology for learning [9], [11], in which they partition the information stream into blocks, and train a model from one piece. We allude to these methodologies as "chunk based" approaches. A troupe of piece based models is utilized to order unlabeled information. These methodologies more often than not keep a fixed-sized outfit, which is consistently refreshed by supplanting a more seasoned model with a recently prepared model. Some lump based systems, for example, [11], can't distinguish novel classes, while others can do as such [9].

Block based methods that can't identify novel classes can't recognize repetitive classes too. This is on the grounds that when a class vanishes for some time, the group in the end disposes of all models prepared with that class. Hence, when the class returns as a repetitive class, none of the models in the troupe can distinguish it. Then again, lump based systems that can identify novel classes, additionally can't distinguish intermittent classes. This is on the grounds that a repetitive class is normally distinguished as a "novel class" by these strategies. Piece based outfit is a group characterization approach, which trains one classifier for every lump, and keeps L (steady) classifiers in the troupe. As it were, the outfit M contains L models {M1, ..., ML}, where every Mi is prepared from one information piece.

V. MEASURES TO IMPROVE THE PERFORMANCE

A. CLASSIFICATION

One of the way to improve the performance of the data stream that has been impacted by the concept drift along with the noise is to adapt a proper classification model. Because the data stream is containing the concept drift already. So if the model is not built for accommodating the changes that are continuously being happening in the data distribution, the model could become obsolete. It can reduce accuracy in classifying. So the classification model must Periodically Re updated. The data that was static before new data block came can be used as starting point. Then the results can be combined which will result in much correct classification as now the model contains latest changes.

B. NOISE REDUCTION

The most general issue that affects the performance is the presence of the Noise in the Data set. Noise is the data points which are included in the data set that are irrelevant. So it can affect the prediction by much difference. So the Noise must be removed either by clustering techniques or outlier detection to improve the efficiency of prediction model.

C. DIMENSIONALITY REDUCTION

In machine learning classification problems, there are often too many factors on the basis of which the final classification is done. These factors are basically variables called features. The higher the number of features, the harder it gets to visualize the training set and then work on it. Sometimes, most of these features are correlated, and hence redundant. This is where dimensionality reduction algorithms come into play. Dimensionality reduction is the process of reducing the number of random variables under consideration, by obtaining a set of principal variables. It can be divided into feature selection and feature extraction.

There are two components of dimensionality reduction:

- **Feature selection:** The technique reduces the input size for the building of Prediction Model and analysis. It involves finding the most important input among all which provides most meaning. Feature Selection gets the only inputs that are affecting the model by a large margin.
- **Feature extraction:** The technique brings down the number of dimensions. Feature extraction derives the new features. It makes the number of attributes in smaller number. The feature extraction technique is much suitable for supervised learning as it improves the speed of the prediction model.

The main difference between Feature Selection and Feature Extraction is that that the Feature Selection keeps the subset of the original features while the Feature Extraction makes a new one.

VI. CONCLUSION AND SCOPE FOR WORK

The purpose of the survey paper was to discuss about the main areas where performance can be affected and at the same time how can be it improved when the data stream contains the concept drift. The paper has been organized regarding the three fundamental zones of research in Concept Drift: chunk ensemble troupes, altering the training set, and group systems.

Though there are many algorithms proposed, many techniques have been stated for the detection of concept drift that affects performance, still the scope for future work exists.

Concept Drift is in certain faculties the incredible glaring issue at hand for AI. The world is consistently changing, however we have a deficiency of methods for comprehension the idea of these progressions as they apply to explicit AI settings. We have a developing collection of advanced techniques for learning with regards to concept drift. There is a need to build up a supporting assemblage of methods for understanding the marvels that these strategies address and in this manner understanding the relative abilities of these strategies even with various articulations of that wonders.

VII. RESULTS AND DISCUSSIONS

Concept shift is firmly identified with concept drift. This happens when a model gained from data examined from one classification should be trained data drawn from another. For instance, a model learned in one dataset may be connected in another dataset, or a model scholarly from client information may be connected to potential clients. For simplicity of piece, this paper concentrates just on this issue of breaking down concept drift, however the methodologies and talk sum up straightforwardly to the similarly imperative issue of concept shift investigation.

Published By:
Blue Eyes Intelligence Engineering
& Sciences Publication (BEIESP)

```
> db
DBSCAN clustering for 999 objects.
Parameters: eps = 0.4, minPts = 4
The clustering contains 24 cluster(s) and 828 noise

   0   1   2   3   4   5   6   7   8   9  10  11  12
828   6  16  13   4  19   7   4   5   4  15   4   4

Available fields: cluster, eps, minPts
>
```

Figure 1. Output1

```
> print(boruta.train)
Boruta performed 10 iterations in 3.336 secs.
 6 attributes confirmed important: X21, X22, X27, X28, X48 and 1 more;
 1 attributes confirmed unimportant: X0.1;
>
```

Figure 2. Output2

```
> boruta.train <- Boruta(X50~.-X0, data = traindat
 1. run of importance source...
 2. run of importance source...
 3. run of importance source...
 4. run of importance source...
 5. run of importance source...
 6. run of importance source...
 7. run of importance source...
 8. run of importance source...
 9. run of importance source...
 10. run of importance source...
After 10 iterations, +3.3 secs:
 confirmed 6 attributes: X21, X22, X27, X28, X48 a
 rejected 1 attribute: X0.1;
 no more attributes left.

>
```

Figure 3. Output3

```
> print(final.boruta)
Boruta performed 10 iterations in 3.336 secs.
 6 attributes confirmed important: X21, X22, X27, X28, X48 and 1 more;
 1 attributes confirmed unimportant: X0.1;
> getSelectedAttributes(final.boruta, withTentative = F)
[1] "X21" "X77" "X28" "X27" "X48" "X22"
> boruta.df <- attStats(final.boruta)
> class(boruta.df)
[1] "data.frame"
> print(boruta.df)
       meanImp   medianImp     minImp    maxImp normHits  decision
X21  6.2737177  6.17550451  5.0569976  7.485274        1 Confirmed
X77 23.5099295 23.11255955 22.1890602 25.194407        1 Confirmed
X28 32.9290697 32.56756660 31.4763028 35.044511        1 Confirmed
X0.1 0.3906841  0.07003618 -0.5296668  2.076816        0  Rejected
X27 35.7514812 35.59626550 34.5703322 37.326526        1 Confirmed
X48 15.9867238 15.94152351 15.5193518 16.563048        1 Confirmed
X22 18.2439983 18.29773298 17.4807369 19.028692        1 Confirmed
>
```

Figure 4. Output4

ACKNOWLEDGMENT

The project "The Impact of Streaming Data Noise Reduction by Using Chunk Based Ensemble" was made possible because of inestimable inputs from everyone involved, directly or indirectly. I would first like to thank my guide, **Asst. Prof. Sr. Nalini N**, who was highly instrumental in providing not only a required and innovative base for the project but also crucial and constructive inputs that helped make my final product.

REFERENCES

1. https://rbi.org.in/scripts
2. www.theukcardsassociation.org.uk, "Card expenditure statistics 2016".
3. Financial Fraud Action UK's Fraud the Facts 2015.
4. Avinash Ingole, Dr. R. C. Thool, "Credit Card Fraud Detection Using Hidden Markov Model and Its Performance", Volume 3, Issue 6, June 2013.
5. V. Bhusari 1, S. Patil, "Application of Hidden Markov Model in Credit Card Fraud Detection", International Journal of Distributed and Parallel Systems (IJDPS), Volume 2, number 6, November 2011.
6. http://www.statsoft.com/Textbook/Fraud-Detection
7. http://profit.ndtv.com/news/your-money/article-financial-fraud-on-the -rise-warns-minister-of- state-1675022
8. [2013] Erkin et al. "Privacy-preserving distributed clustering'' EURASIP Journal on Information Security Erkin et al.; licensee Springer http://jis.eurasipjournals.com/content/2013/1/4.
9. Bolton, Richard J. & Hand, David J, "Statistical Fraud Detection: A Review", Statistical Science, Volume 10, number 3, pp 235, 2002.
10. Sushmito Ghosh and Douglas L. Reilly Nestor, "Credit Card Fraud Detection with a Neural-Network," Inc. Proceedings of the Twenty-Seventh Annual Hawaii International Conference on System Sciences, pp.621-630, 1994.
11. Peipei Li , Xuegang Hu , Qianhui Liang , Yunjun Gao, Concept Drifting Detection on Noisy Streaming Data in Random Ensemble Decision Trees, Proceedings of the 6th International Conference on Machine Learning and Data Mining in Pattern Recognition, July 23-25, 2009, Leipzig, Germany
12. Brzezinski, D. and Stefanowski, J. (2014). Reacting to Different Types of Concept Drift: The Accuracy Updated Ensemble Algorithm. IEEE Transactions on Neural Networks and Learning Systems, 25(1), pp.81-94.
13. P. S. Cowpertwait and A. V. Metcalfe, Introductory time series with R. Springer, 2009.
14. J.-Z. Wang, J.-J. Wang, Z.-G. Zhang, and S.-P. Guo, "Forecasting stock indices with back propagation neural network," Expert Systems With Aplications, vol. 38, no. 11, pp. 14 346–14 355, 2011.
15. Masud, M., Chen, Q., Khan, L., Aggarwal, C., Gao, J., Han, J., Srivastava, A. and Oza, N. (2013). Classification and Adaptive Novel Class Detection of Feature-Evolving Data Streams. *IEEE Transactions on Knowledge and Data Engineering*, 25(7), pp.1484-1497.
16. LI, N., GUO, G. and CHEN, L. (2013). Concept drift detection method with limited amount of labeled data. Journal of Computer Applications, 32(8), pp.2176-2181.
17. LEE, J. and LEE, Y. (2011). Concept Drift Detection for Evolving Stream Data. *IEICE Transactions on Information and Systems*, E94-D(11), pp.2288-2292.
18. Krawczyk, B., Stefanowski, J. and Wozniak, M. (2015). Data stream classification and big data analytics. *Neurocomputing*, 150, pp.238-239.
19. LEE, J. and LEE, Y. (2011). Concept Drift Detection for Evolving Stream Data. *IEICE Transactions on Information and Systems*, E94-D(11), pp.2288-2292.
20. LI, N., GUO, G. and CHEN, L. (2013). Concept drift detection method with limited amount of labeled data. *Journal of Computer Applications*, 32(8), pp.2176-2181.
21. Yu, S., Abraham, Z., Wang, H., Shah, M., Wei, Y. and Príncipe, J. (2019). Concept drift detection and adaptation with hierarchical hypothesis testing. *Journal of the Franklin Institute*, 356(5), pp.3187-3215.
22. Krawczyk, B., Stefanowski, J. and Wozniak, M. (2015). Data stream classification and big data analytics. *Neurocomputing*, 150, pp.238-239.
23. J. Gama, I. Zliobait ˙ e, A. Bifet, M. Pechenizkiy, and A. Bouchachia, "A ˙ survey on concept drift adaptation," ACM Computing Surveys (CSUR), vol. 46, no. 4, p. 44, 2014.
24. M. Friedman, "A comparison of alternative tests of significance for the problem of m rankings," The Annals of Mathematical Statistics, vol. 11, no. 1, pp. 86–92, 1940.
25. I. Katakis, G. Tsoumakas, and I. Vlahavas, "An ensemble of classifiers for coping with recurring contexts in data streams," in Proceedings of the 2008 conference on ECAI 2008: 18th European Conference on Artificial Intelligence. IOS Press, 2008, pp. 763–764.