A Review on Streaming Data and Decision Tree Classifiers for Non-stationary Data

Satyajit S. Uparkar[1],Dr. Ujwal A. Lanjewar [2]

[1] Research Scholar , Inter Institutional Computer Centre, RTM Nagpur University, Nagpur, India

[2] Professor, Department of MCA, ICSR-VMV Commerce JMT Arts and JJP Science College Nagpur, India

[1] ssuparkar@yahoo.co.in,[2]ualanjewar@gmail.com

ABSTRACT

In streaming data, most of time the past data, does not make any sense for the prediction of current and future analysis of an event due to many circumstances and noise. This result into a poor prediction and less accurate model which happens because, the classifier is trained to work on the past data. In this paper, we are presenting the basic skeleton and existing methods in non-stationary type of data. This paper also discuss the weakness of current decision tree classifier i.e. problem associated with the searching of split point, memory size and running time complexity associated with classifier for non-stationary data as learning with small dataset and huge dataset is completely different. This paper gives an emphasis in the domain of speculations, statistical reasoning and forecasting to explore and understand the problems associated with the prediction model when it applies to a non-stationary data.

Keywords

Data-driven-decision, Non-stationary data, Decision tree classifier

Article Received: 10 August 2020, Revised: 25 October 2020, Accepted: 18 November 2020

Introduction

The data is growing in an exponential manner and at a countless rate, which is coming up from a collection of equipments, networks, stock transactions, banking transactions and cell phones etc [1]. According to the survey in 2018, it is found that there is around 25ZB i.e. around 25 billion TB of data worldwide. Based on the calculation and the percentage growth, we can expect the growth in the data in the year 2025 would be around 170ZB [2]. As such there is no standard definition of data stream, it changes from author to author in their explanation, but if we talk about data stream or non-stationary data, it is nothing but the data must be received or generated by the system or a user in a logical order and in the continuous way. The another way to express the definition of streaming data is, when we receive the data from multiple sources i.e. from the heterogeneous sources is referred to as non-stationary data. This non-stationary data contains a very important information provided it can be extracted in a timely manner, before it losses it qualitative and quantitative information [3]. Various machine learning methodologies had been proposed which are adaptable in nature and having capability to handle non-stationary data, is used to find the information or trends associated with the online data; this is widely known as stream mining [4]. Due to the high velocity and continuous change in the data pattern with respect to its size, it becomes very difficult to handle the streaming data and comes up with many challenges for real time interpretation, forecasting and prediction.

In this research paper, we had provided a detailed study of the non-stationary data and its mining with a emphasis on challenges which are associated with the mining of non-stationary data. This study is concerned with the supervised non-stationary techniques. This study highlights the limitations of existing decision tree classifiers with respect to the non-stationary data.

LITERATURE REVIEW

The researchers working in the field of stream mining and machine learning very well knows that the non-stationary data grows exponentially and having very high velocity compare to the stationary datasets. The problem associated with the non-stationary data is the concept drift which makes it difficult to train the prediction model or classifier to work with the continuous stream of heterogeneous data. Hence, it is very much necessary for the algorithms to get adaptive with the resources availability and various other factors. As non-stationary data is having evolving

nature we had done detailed survey on present non-stationary data prediction algorithms like rule based, hybrid or ensemble, nearest neighbor, statistical and tree structured [5]. The below mentioned table i.e. Table No. 1, gives complete description of the algorithms.

Table 1 : Comparison of various non-stationary data Classification techniques

Sr. No.	Non- stationary Data Classification Techniques	Publication year	Key Highlights	Classification group
1	Incremental Tree Induction [6]	1997	The proposed algorithm requires a huge amount of memory and is the only reason not applicable for massive non-stationary data stream.	Based on Tree methods
2	Very Fast Decision Tree Learner [7]	2000	The proposed method will be useful, if there is same amount of correct trees compared with the traditional system and computing resources is same. It consumes less memory and can forecast in an online environment. In this technique, Hoeffding method is used for doing the calculation of output with respect to the conventional learner.	Based on Tree methods
3	Ensemble Based Technique for non-stationary data [8]	2001	This paper proposed an technique of anytime learning for the given problems of any type (it may be huge or small). This method works with independently with respect to the said classifier.	Based on Ensemble Method
4	OD-Classification on non-stationary data [9]	2004	In this, dynamic approach is used for both training and testing purpose of classification of non-stationary data-streams. The classifier is trained in live scenario over the non-stationary datasets, here goal is to develop a classification model which can easily adapt the change in the live data.	Based on Rule
5	VHT- Vertical Implementation of Hoeffding tree [10]	2013	It uses the method of partitioning of the non-stationary data vertically and performs parallel computations.	Based on Tree methods
6	Non-stationary mining classifiers [11]	2013	In this paper author proposed an algorithm which is based on VFDT method.	Based on tree method

7	SimC [12]	2014	It easily captures the change in non-stationary data and builds the model for immediate prediction.	Based on rule
8	MNS-Learning using Decision Trees [13]	2014	It uses a decision tree for the learning of markov structure.	Based on Tree Method
9	OSC-based Incremental semi-supervised learning [14]	2015	This method is useful when the training non-stationary data is very limited, i.e. very less number of labels are available.	Based on rule
10	Classifier based on Clustering approach [15]	2015	It uses the distance metrics and uses kernel method of clustering.	Based on rule
11	Dynamic Classification model [16]	2015	The proposed model uses support vector machine and based on incremental method.	Based on rule
12	Plug-and-Play Dual-Tree Algorithm Runtime Analysis [17]	2015	The proposed method explores the relationship between present data and past data in a rapid manner. It easily identifies the change in the incoming non-stationary data with less time and space complexity.	Based on tree method and nearest neighbor.
13	Adaptive Decision Trees for UHF and LD [18]	2017	This paper proposed a novel algorithm for handling ultra high feature of non-stationary data.	Based on tree method
14	To tune or not to tune the number of trees in random forest [19]	2018	This paper proposed a method for fast prediction of non-stationary data by using the less number of resources. It consumes less amount of memory compare to the algorithm mentioned. It is adaptive with the change in the data distribution	Based on tree method
15	FS- based DT [20]	2019	It is the most improved algorithm which uses a greedy top-down tree induction method for handling the non-stationary data and gives a good accuracy.	Based on tree method

CONCLUSION & FUTURE SCOPE

There are many classification algorithms available which works good with stationary datasets but they are mostly incompatible to work with non-stationary data due to less amount of memory and takes very large time for accessing the data. This paper presents the detailed study on various algorithms which are designed for handling non-stationary data. We had explored the minute details of each and every classifier which were developed since 1997-2019 and what classification techniques they used. In our future research work, we will be proposing a new scalable decision tree classifier which will overcome the limitation of present classifiers and

will be based on histogram method for classification of huge datasets and non-stationary data.

REFERENCES

[1] Aggarwal C, 'Data streams: models and algorithms', Springer, Berlin, Vol.31, 2007.

[2] Reinsal D, Gantz J, Rydning J., 'Data age 2025: 'The evolution of data to life-critical don't focus on the data that's big'.

[3] Technical reports on IDC. https://www.seagate.com/www-content/our-story/trends/files/Seagate-WP-DataAge2025-March-2017.pdf. Accessed 14 Oct 2019.

[4] Ditzler G, Roveri M, Alippi C, Polikar R., 'Learning in non-stationary environments: a survey', IEEE Comput Intell Magn 10:12-25, 2015.

[5] Zliobaite I, Bifet A, Pfahringer B, Holmes G., 'Active Learning with drifting streaming data', IEEE Trans Neural Netw Learn Syst 25(1): 27-39, 2014.

[6] Prasad, Bakshi R., Sonali A., 'Stream data mining: platforms, algorithms, performance evaluators and research trends', International journal of database theory and application, 9.9, 201-218, 2016.

[7] Utgoff, Paul E., Neil C. Berkman, and Jeffery A. Clouse. 'Decision tree induction based on efficient tree restructuring', Machine Learning -29.1, 5-44, 1997.

[8] Domingos, Pedro, and Geoff Hulten. 'Mining high-speed data streams', Proceedings of the sixth ACM SIGKDD, International conference on Knowledge discovery and data mining, ACM, 2000.

[9] Street, W. Nick, and YongSeog Kim, 'A streaming ensemble algorithm (SEA) for large-scale classification, Proceedings of the seventh ACM SIGKDD international conference on Knowledge discovery and data mining', ACM, 2001.

[10] Aggarwal, CharuC., et al. 'On demand classification of data streams, Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining', ACM, 2004.

[11] Wang, Lei, Hong-Bing Ji, and Yu Jin. 'Fuzzy Passive–Aggressive classification: A robust and efficient algorithm for online classification problems', Information Sciences- 220, 46-63, 2013.

[12] Simon Fong, Yang Zhang, Jinan Fiaidhi, Osama Mohammed, Sabah Mohammed, 'Evaluation of stream mining classifiers for real-time clinical decision support system: A case study of blood glucose prediction in diabetes therapy', Hindawi publishing corporation, Vol. 2013, doi.org//10.1155/2013/274193.

[13] Brzezinski, Dariusz, and Jerzy Stefanowski, 'Prequential AUC for classifier evaluation and drift detection in evolving data streams', International Workshop on New Frontiers in Mining Complex Patterns. Springer International Publishing, 2014.

[14] Daniel L, Jesse D, 'Improving Markov Netwok Structure Learning Using Decision Trees', journal of machine learning research, 15, 501-532, 2014.

[15] Loo, HuiRu, and Muhammad N. Marsono, 'Online data stream classification with incremental semi-supervised learning', Proceedings of the Second ACM IKDD Conference on Data Sciences. ACM, 2015.

[16] Jdrzejowicz, Joanna, and Piotr Jdrzejowicz., 'Distance-based ensemble online classifier with kernel clustering, Intelligent Decision Technologies', Springer International Publishing, 279-289, 2015.

[17] Krawczyk, Bartosz, and Michał Woźniak, 'One-class classifiers with incremental learning and forgetting for data streams with concept drift', Soft Computing -19.12, 3387-3400, 2015.

[18] Ryan R. Curtin, Dongryeo Lee, William B. March, Parikshit Ram, 'Plug-and-Play Dual-Tree Algorithm Runtime Analysis', Journal of machine learning research, 16, 3269-3297, 2015.

[19] Weiwei Liu, Ivor W. Tsang, 'Making Decision Trees Feasible in Ultrahigh Feature

and Label Dimensions', journal of machine learning research, 18,1-36, 2017.

[20] Philipp Probst, Anne- Laure B, 'To tune or not to tune the number of trees in random forest', journal of machine learning research ,18,1-18, 2018.

[21] Salvatore R, 'Complete search for feature selection in decision trees', journal of machine learning research 20,1-34, 2019.