November-December 2019
ISSN: 0193-4120 Page No. 4493 - 4502

A General Framework for Streaming Data Analytics

D.Christy Sujatha[1], Dr.J.Gnana Jayanthi[2]
[1] Research Scholar, PG and Research Department of Computer Science, Rajah Serfoji Government College, Thanjavur, Tamil Nadu, India, christy_se@pmu.edu
[2] Assistant Professor, PG and Research Department of Computer Science, Rajah Serfoji Government College,Thanjavur,Tamil Nadu, India ,jgnanajayanthi@gmail.com

Article Info
Volume 81
Page Number: 4493 - 4502
Publication Issue:
November-December 2019

Abstract

Real time applications encompasses with streaming data which comes in huge volume and at high speed. In recent years, data has become one of the most important features of any organization for its progression. Most of the organizations are turning into accomplishing real time data analytics to get the speedy outcome using existing digital data generated every day. But building a predictive model for any real time data is a more challenging one which attracts the attention of most of the Researchers, Academicians and Industrial Persons. The objective of this study is to design a common framework to show various pathways in handling streaming data which can be used to identify the challenges and solving techniques by the streaming research community. This general framework illustrates several existing streaming data attaining sources, challenges and classification algorithms in handling streaming data. It also comprises some of the available platforms, performance metrics and applications using streaming data.

Article History
Article Received: *5 March 2019*
Revised: *18 May 2019*
Accepted: *24 September 2019*
Publication: *23 December 2019*

Keywords: *Challenges, Classification algorithm, Data analytics, Real time applications, Streaming data framework, Streaming data platforms, Performance Metrics*

1 INTRODUCTION

Streaming data analytics or real time analytics is an exclusive type of Big Data analytics, where real time data elements are processed and significant insights will be delivered instantaneously in their arriving order[1]. Hence most of the organizations are turning towards streaming data analytics to identify the probable outcome immediately [2]. Due to continuous, unbounded and rapid speed of streaming data, there is not sufficient amount of storage to store the entire data and multiple scanning of data stream is not possible and hence the traditional analytics algorithms are not suitable to process data stream [3]. This unbounded storage and rapid speed of data stream causes, the growing need of novel analytical techniques for streaming data [4].

A framework is a model, designed to outline an analyst's logical thinking in a systematic manner which guides and facilitates understanding of the problem [5]. In this survey paper several approaches handled by the data stream researchers are analyzed and a general framework is designed to illustrate various pathways in handling streaming data which can be used to identify the challenges and solving techniques by the streaming research community. This general framework starts with frequently used ways to collect streaming data and ends with performance metrics of the built model. It also covers some of the research issues of streaming data,

November-December 2019
ISSN: 0193-4120 Page No. 4493 - 4502

classification techniques used by the researchers to solve the problems of data stream. Besides that it also figures out some of the existing open source analytical tools to process data stream. Because if the captured data is not processed in a certain amount of time it's value will be lost and hence data stream needs additional attention in usage of appropriate tools [6]. Finally a comparative study is preformed for various applications of streaming data analytics along with it's data sources generated and platforms.

This survey paper commences with Introduction in section 1 and in section 2 a general framework is illustrated with neat diagrammatical representation. Various components of the framework like Data sources, Key issues, Classification algorithms, Performance metrics and tools in handling streaming data are discussed. In Section 3 a relative study of several applications using streaming data is illustrated. Finally section 4 ends with conclusion.

2 General Framework For Streaming Data Analytics

A methodical study and review is performed by considering several research papers published by several streaming data researchers and a general Frame Work is designed. This general framework is shown in Fig 1. that covers different aspects of streaming data which is helpful in understanding streaming data for further research work. The Frame work consists of several Streaming Data sources, Issues, Technical solutions, Performance Metrics, Platforms to handle streaming data.

2.1 Data Source

Data streaming is a process of sending data records continuously. From the survey it is found that most of the streaming data researchers used the data from social network [7,8], sensors [9],[10],[11], log files generated using web applications[12],[13], Mobile Applications[14] , information from social networks [15],[16], synthetic data (or) data generated from the artificial environment [17],[18],[19].

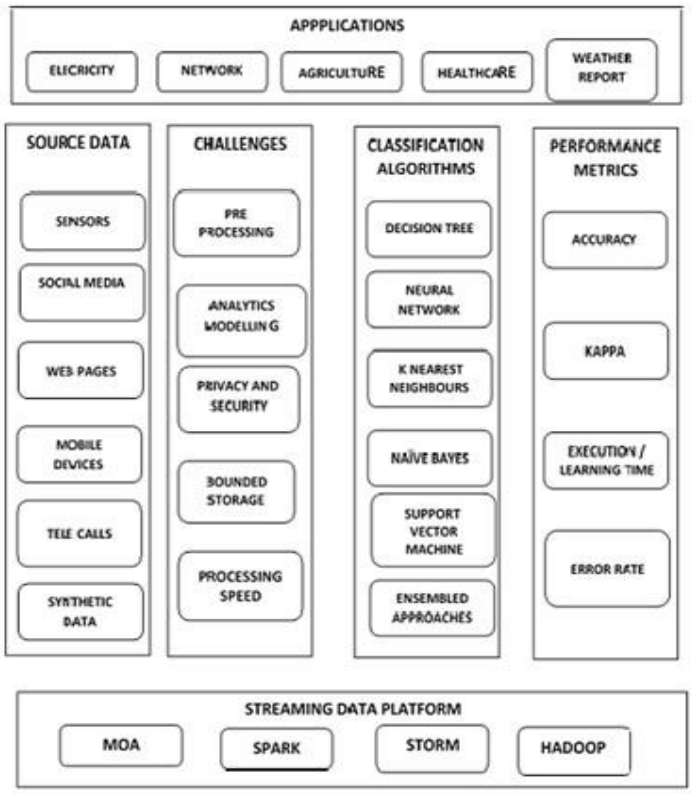

Fig 1. General Frame for Streaming Data Analytics

2.2 Challenges Of Real Time Streaming Data

This section involves various key challenges resolved by several Researchers , Academicians and Industrialists using streaming data.

2.2.1 Pre processing

The collected streaming data from various sources have to be pre-processed effectively before passing the data to the analytics phase. The raw data may have NULL values, duplicate values and erroneous values. If this uncleaned data is sent to the analytics phase we may not get the accurate result in prediction. Hence suitable Machine learning algorithms are needed in pre-processing techniques in order to remove NULL values, incomplete data, erroneous and outlier data [20]. While cleaning streaming data one should be aware not to remove any useful information which is an another key challenge. The author Jayaram Hariharakrishnan et al., suggested it is also possible to automatically generate the right metadata to define what data is

November-December 2019
ISSN: 0193-4120 Page No. 4493 - 4502

recorded, how it is recorded and measured, location of the data generated which will be used for further analytics phase [21].

2.2.2 Analytics Modelling

Analytics modelling is used to empower the decision making of any organization to progress in a success way. Using this analytics model, the hidden information that hide in raw data is bringing out either in the form of graphical report or data visualization. Hence it is another great challenge to build the analytics model based on the requirement of the end user which may be of Descriptive, Prescriptive or Predictive [22]. Most of the work done so far were offline based or static based where the entire data is stored and processed later. Mahnoosh Kholghi et al., proposed an analytical tool to get streaming data through on line which hits from various resources at every second with higher speed [23]. Privacy and Security is another sensitive challenge with streaming data which deals with the basic rights of an organization or an individual. An organization *or* an individual has to decide what data can be shared with third parties and what data to be maintained with privacy and security [24]. In order to enhance the business, the personal details of an individual is collected by the business organization through web page to infer new knowledge about the customer regarding their buying ability, product demand and sales. The information collecting organization may store the customer's personal data in their private storage or cloud storage. If no proper security is provided in the cloud storage the private data of a customer may be stolen by a third party. It is a big challenging factor to maintain the personal data of an organization or individual [25]. So it is necessary to implement appropriate privacy and security to the big data, otherwise it may lead to the failure of the technology implementation and some unpleasant results.

2.2.4 Bounded Storage

It is not possible to store the entire streaming data set in main memory or on disk and the data set is scanned only once [26]. So the traditional storage techniques cannot be used for streaming data storage. Hence in some cases the summaries of data stream can be considered and stored. Some of the authors developed new summarisation technique for the stream to produce approximate solutions [27].

2.2.5 Processing speed

The processing speed of any algorithm depends on the arrival time and the incoming data and finding correlation among the incoming data streams and processing algorithm is another key issue in streaming data. Because the data distribution in streaming data changes over time and the built predictive model is not suitable for the upcoming data stream. The authors Supun Kamburugamuve and Michael Hahsler proposed predictive algorithm which is adaptable to the rapid speed of streaming data [28],[29].

2.3 Classification Algorithms in handling streaming data

This section illustrates several classification algorithm used by the existing researchers to propose a solution in handling streaming data.

2.3.1 Decision Tree

Most of the authors [30],[31],[32] used decision tree to handle streaming data in their research work. A decision tree is a graphical illustration of probable solutions to a decision based on some condition which starts from the root node and branches down to the number of solutions like a tree. Decision trees are non-parametric, sensitive to the outliers. It learns by recursively replaces the leaves by applying heuristics measures.

2.3.2 Artificial Neural Network(ANN)

Tomoyasu Takata [33] and Zeineb Hammami et al.[34], used artificial neural network in their

research work. Artificial Neural Network is an interconnected collection of nodes, similar to the network of neurons in human brain. Each node represents an artificial neuron and an arrow represents a connection from the output of one artificial neuron to the input of another. ANN is a non linear and non parametric model and it learns by applying gradient descent procedure on newly arriving streaming data.

2.3.3 K Nearest Neighbors (KNN)

KNN or K - Nearest Neighbours, is one of the simplest Supervised Machine Learning algorithm used for both classification and regression predictive problems. KNN is a non parametric technique, used for classification and regression. It uses *k*, which is the number of its nearest neighbours, to classify data to its group membership. Since it stores the data only at training time, it is also denoted by Lazy leaners .[35],[36].

2.3.4 Support Vector Machine (SVM)

Isah A Lawala,Salihu et al [37] , Pranamita Nanda,B et al. [38]

used Support Vector Machine which can perform both linear and non linear data set. Linear SVMs can be trained more quickly, but they are less accurate than non-linear approaches.

2.3.5 Naïve Bayes(NB)

It is a parametric, incremental classification technique based on Bayes' Theorem with an assumption of independence among predictors. The major drawback is it is also known as a bad estimator and finding fully independent predictors is very difficult to achieve in real life[39].

2.3.6 Ensemble Approach Algorithms

The authors Leandro L. Minku et al., [40], Joao Gama [41] and Parneeta Sidhu, M used Ensemble approaches by combining the above mentioned base learner algorithms in order to improve the accuracy of the prediction.

In Table 1 the basic classification algorithms are presented with its Merits and Demerits of the algorithms.

TABLE 1
Comparison of Classification Algorithms For Streaming Data

Description	Merits	Demerits	Ref
Decision Tree is a graphical representation of possible solutions to a decision based on certain condition.	Non-parametric, Distribution free, Robust to the outliers and irrelevant attributes., High degree of interpretability, Automatic Feature Selection, Handle both Numerical and Categorical data ,Training cost is very less in terms of logarithmic.	Leads to overfitting and Class imbalance Problems.,	[30], [31],[32
Artificial Neural Network is an interconnected group of nodes, infer unseen relationships on unseen data	Non linear , Non parametric Can infer hidden relationship in the data by training the model	Hard to interpret, visualise and understand the weights of a neural network and consumes a large amount of time for training the model. Not accurate result for classification	[33] [34]
K Nearest Neighbours uses *k*, the number of its nearest neighbours, to classify data to its group membership	Simple, cheap, easy to understand and implement and fast in pattern recognition.	Also called as lazy learner, Choosing k value is difficult., Poor performance when dimension grows, Computation cost grows when data size grows	[35],[36]
Support Vector Machine, analyze data used for classification and regression analysis into two groups.	Can perform both linear and non linear classification,, Compact representation of historical data in terms of support vectors.	Runs slow, Computationally expensive, Performs , poor performance in multi classification	[37],[38]
Naïve Bayes based on Bayes' Theorem with an assumption of independence among predictors.	Less training time and less training data, can handle both discrete and continuous data.	Also called as bad estimator. it is difficult to get the completely independent predictors.	[39]
Ensemble Learning trains multiple learners to solve the same problem.	Improves accuracy of the prediction	Consumes more memory, more cost , to train and maintain more than one classifier.	[40],[41],[42

2.4 Performance Metrics

Most of the streaming data researchers used Prediction Accuracy ,Kappa, Error rate , Memory Usage, Recall rate , Scalability [43],[44],[45],[46] in order to measure their proposed algorithm and to compare with the existing algorithm.

2.5. Streaming Data Platforms

This section illustrates some of the tools to process on line data and it's comparative study which has been tabulated in Table 2.

2. 5.1 Hadoop Map Reduce

Hadoop MapReduce is a framework for processing large amount of data which has Hadoop Distributed File System(HDFS) to store more volume of data. It also provides fault tolerance by keeping replication of data[47]. The Master job tracker receives the incoming task and allotted the job to the slave tasks. Since it needs all data at the beginning, it can only be used for stored batch process. It takes

November-December 2019
ISSN: 0193-4120 Page No. 4493 - 4502

extra time to read and write any operation since it keeps the intermediate result in hard disk.

2.5.2 Apache Spark

Apache Spark can perform both stored data as well as on line data [49]. The intermediate result is stored using Resilient Distributed Dataset (RDD) abstraction [50] which increases processing performance of Spark [51]. It supports both Clustering and Classification algorithms to extract features from the data set.

2.5.3 Apache Storm

Apache Storm performs the task in parallel to speed up the process [53]. It uses HDFS file system for the distributed operations. The workflow is defined with the help of Directed Acyclic Graphs (DAG's) by using Marathon framework. [54].

2.5.4 Massive On line Analytics (MOA) Frame Work

MOA (Massive On line Analysis) is especially developed for data stream that includes algorithms, evaluation tools and artificial data producers. Most of the authors used MOA frame work due to it's ease of usage[55].

3. Applications of Streaming Data Analytics

Hassan Nazeer et al.[56], proposed a real-time text processing pipeline using open-source big data tools which minimize the latency to perform sentiment analytics for twitter data gathered through Twitter streaming API. The author used Apache Kafka for data ingestion system and Apache Cassandra for persistent distributed storage.

TABLE 2
COMPARISON OF STREAMING DATA PLATFORMS

	HADOOP MAPREDUCE	SPARK	STORM	MOA
Architecture	Master Slave	Master Slave	Peer	---
Data Processing Engine	Supports only Batch Processing	Supports both batch and stream processing	Supports Stream process	Supports Stream process
Processing Speed	Much slower than others due to map and reduce task	100 times faster than Hadoop due to In memory process	Can process thousands messages per second on cluster	Slower compared with others
Intermediate storage	Intermediate result is stored in memory	Intermediate result is stored in RDD	No intermediate storage	–
Programming Language used	Java, C#, Python and R	*Java, Scala, Python, and* **R.**	*Java and Scala.*	Java
Scheduler	It needs External Job scheduler like oozie	It has it's own scheduler	It has it's own scheduler	–
Response Time	Minutes to hours	Seconds	Milli seconds	Slower than other
Development Cost	Coding available only for batch processing	We can use same coding for stream processing as well as batch processing	We cannot use the same code for both processing	A free Open Source Project
Installation	Easy to install	Easy to install	Not easy to install	Easy to install
REFERENCES	**[11, 50]**	**[20, 35]**	**[19, 23]**	**[16,17,18,22,24, 37,45,46,48]**

Adnan Akbar et al.[57], proposed an architecture to predict the complex pattern of real time events generate from IoT called Complex Event Processing (CEP). The Proposed prediction model utilizes moving window of data for training the model using adaptive prediction algorithm called Adaptive Moving Window Regression (AMWR) for dynamic IoT data.

Balaji,S et al [58] Proposed Artemis Cloud framework, an extension of Artimis Big data on line health analytics frame work which was deployed in NICU at SickKids Hospital in Toronto in August 2009. The proposed framework is a remote real time patient monitoring using low resource settings, and used android device for data visualization targeted for a small unit NICU/PICU set up in India. The author derived heart rate value from electrocardiogram signal and pulse rate from plethysmography waves where each peak corresponds to one heart beat. Thanga prasad.S et al., [59] Proposed a framework for the benefit of

November-December 2019
ISSN: 0193-4120 Page No. 4493 - 4502

Diabetic patients. The author used Electronic demo tool for data collection and Hadoop map reduce tool for data processing. The author proposed a web interface to analyse patients health by collecting HER records, stored in HBASE and processed by Hadoop frame work. Data is accessed by HDFS and queries executed by HIVE.

Basil Shaik et al., [60] proposed a web interface to analyze health records using Hadoop environment and M.R.Bendre et al., [61] developed a predictive model using Linear Regression algorithm, a supervised machine learning technique to predict about Temperature and Rainfall. They implemented Hadoop Map reduce programming environment to find the mean values and to increase the speed of execution. Google File System(GFS) was used to distribute the data over network. The model predicts rainfall and temperature values for the year 2013 and also compared actual and predicted values to minimize the error.

K. L. Ponce-Guevara et al.[62] Proposed a software tool / interface GreenFarm-DM developed a predictive model to predict about soil moisture and to optimize water usage. Suhas Athani et al,[63], proposed a project work in which he collects data from soil moisture sensor connected to a Arduino and the output is connected with android application using wifi shield. The collected data is processed by neural network algorithm for the correction factors . these output values are shown to the farmers through their mobile phones.

RabiaLatif, HaiderAbbasv et al., [64] proposed a model to classify the traffic in the network using decision tree and he implemented the research using MOA framework Bobin K. Sunny et al., [65] Proposed a Predicting and Recommending TV channel to the viewers based on their viewers past history profile, the system identifies the most appropriate channel to the user. The author used the combination of Logic Regression with Stochastic Gradient Descent for training process and self adaption techniques to set the tuning factor for the new user and they used Spark streaming and Lambda Architecture for the analytics process. Implementation of Spark streaming and the regression algorithm is not clearly described. For the new user, the system recommends less accurate prediction. More memory and computation complexity due to the storage of millions of viewers individual profile for the channel .recommendation

Kazem Fathi et al [66] combined decision tree and KNN to find the intrusion and Kalpesh Adhatrao et al., [67] used decision tree to predict the students performance, Simon Fong, et al., [68] used MOA framework and Decision tree algorithm to analyze ECG and EEG signals.

The above discussed applications are compared and tabulated in *Table 3* by considering Tools, Algorithms and source data taken.

TABLE 3
COMPARISON OF STREAMING DATA APPLICATIONS

Description	Tool Used	ML algorithm	Source Data Size / Type/	Ref
Text Analytics for Twitter data	APACHE SPARK	Naive Bayes classifier	Data is loaded from Twitter	[56]
To Predict the Complex Events for IoT data by combining ML with CEP	APACHE SPARK APCHE KAFKA	Adaptive Moving Window Regression (AMWR)	Traffic data provided by city of Madrid.	[57]
An on line health analytics frame work.	Android device, Cloud Services	---	Health records Electrocardiogram signal, Plethysmography waves.	[58]
To predict and Classify the types of Diabetic Mellitus	Hadoop Map Reduce.	Predictive Investigation algorithm	Diabetics data	[59]
Web Interface to analyse Health records	HADOOP HBASE, HIVE	---	HER records	[60]
To predict the temperature and rain fall	Hadoop Map Reduce ,Google File System	Linear Regression Algorithm	Temperature, humidity and rainfall data of Krishi Vidyapeeth Rahuri	[61]
To predict about soil moisture to optimize water usage.	Green Farm with Wireless sensor Network	C4.5 algorithm (Decision tree)	Data collected from a wireless sensor network (WSN)	[62]
To monitor soil moisture and predict the rainfall for the welfare of the farmers.	Arduino	Neural network algorithm	Moisture , PH content and salinity content taken from soil moisture sensor	[63]
To Classify the Network Traffic	MOA	EVFDT (Decision Tree)	---	[64]
To Predict and Recommend TV channel	SPARK	Logic Regression	Real time click stream data	[65]
Intrusion Detection System	WEKA	Decision Tree and KNN	University of California, Irvine Knowledge Discovery and Data Mining	[66]
To predict Students performance	Rapid MIner	Decision Tree (ID3, C4.5)	First Year Students Information System	[67]
ECG , EEG signals Analytics	MOA	OVFDT (Decision Tree)	Life science data sets from UCI Repository	[68]

November-December 2019
ISSN: 0193-4120 Page No. 4493 - 4502

5 4. Conclusion

In this survey paper a general framework for streaming data is designed to illustrate various pathways in handling streaming data which can be used to identify the challenges and solving techniques by the streaming research community. This general framework comprises of several data source collection methods and focusses some of the research issues, classification algorithms used by the researchers to solve the problems of data stream. Besides that it also figures out some of the existing open source analytical tools to process data stream. Finally various applications of streaming data analytics is analyzed and reviewed with a comparative study.

REFERENCES

[1] Z. Milosevic, W. Chen, A. Berry, and F. A. Rabhi, "An event-based model to support distributed real-time analytics: finance case study," presented at the EDOC, 2015.

[2] Babak Yadranjiaghdam, A Survey on Real-Time Big Data Analytics: Applications and Tools, Conference: 2016 International Conference on Computational Science and Computational Intelligence (CSCI),December 2016

[3] Yunyue, and Dennis Shasha Zhu, "Statstream: Statistical monitoring of thousands of data streams in real time," in In Proceedings of the 28th international conference on Very Large Data Bases, pp. 358-369, 2002.

[4] B. Ellis, Real-Time Analytics: Techniques to Analyze and Visualize Streaming Data: Wiley, 2014

[5] Amir Kandomi, Murtaza Haier, Beyond the hype Big data concepts methods and Analytics, In: International Journal of Information Management 0268-4012,2014.

[6] http://opensourceforu.com/2017/09/open-source-tools-you-can-use-to-handle-big-data

[7] Big Data A New World of Opportunities , NESSI White Paper, December ,2012.

[8] Vaddadi Vasudha Rani , Sandhya Rani, Twitter Streaming and Analysis through R, In : Indian Journal of Science and Technology, Vol 9(45),2016

[9] Sandeep Singh Sandha, Complex EventProcessing of Health Data in Real-time to Predict Heart Failure Risk and Stress , In: arXiv :1707. 04364v1 [Cs.Cy] ,2017

[10] Farrukh Aslam Kahan , " A continuous Change Detection Mechanism to Identify Anomalies in ECG signals for WBAN Based healthcare Environment" , in Special Section on Security analytics and Intelligent for cyber physical system, IEEE Access , 2017

[11] Anjit Ukil ,Soma Banyoapdhyay , " IoT Health Care Analytics : The Importance of Anomaly detection " : In 2016 IEEE 30th Conference on Advanced Information Networking and application .

[12] Sahar Yassine ," A framework for Learning Analytics in Moodle for Assessing Course outcome", 2016 IEEE Global Engineering Education conference (EDUCON)

[13] Rianne Conjin, Chris Sniders, Ad Kleingeld , and Uwe Matzat , " Predicting Student Performance from LMS Data ; A comparison of 17 Blended Courses Using Moodle LMS " , IEEE Transaction on Learning Technologies , Vol 10 No. 1 , March 2017

[14] Boshi Li, Rita Kuo, Maiga Chang, Kristin Garn ," reward Points Calculation based on Sequential Pattern Analysys in an Educational Mobile App",

[15] M.Sridevi, B.R Arunkumar, "Social Network Analysis And Its Applications -A Review From Business Perspective" International Journal of Informative & Futuristic Research (IJIFR) Volume - 2, Issue - 9, May 2015 21st Edition, Page No: 3006-3013

[16] Roberto Interdonato, Andrea Tagarelli. "Ranking Silent Nodes in Information Networks: a quantitative approach and applications". Elsevier Publication, January, 2014.

[17] Parneeta Sidhu, M. P. S. Bhatia, An online ensembles approach for handling concept drift in data streams: diversified online ensembles detection , In : Springer DOI 10.1007/s13042-015-0366-1 ,2015

[18] Ignsio Cano, Mohammad Raza kha , ASML: Automatic Streaming Machine Learningn, In : 1https://github.com/nachocano/asml, 2015

[19] Isvani Frías-Blanco, Alberto , Fast Adaptive Stacking of Ensembles , In : ACM ISBN 978-1-

November-December 2019
ISSN: 0193-4120 Page No. 4493 - 4502

4503-3739-7/16/04, 2016

[20] Challenges and Opportunities with Big Data, A community white paper developed by leading researchers across the united states.

[21] Jayaram Hariharakrishnan, Mohanavalli,.S, Srividya, Sundhara Kumar, K.B, Survey of Pre-processing Techniques for Mining Big Data, In : IEEE International Conference on Computer, Communication, and Signal Processing ICCCSP, 2017

[22] Michael Hahsler, Matthew Bolanos, Introduction to stream: An Extensible Framework for Data Stream Clustering Research with R,2017

[23] Mahnoosh Kholghi , Mohammadreza Keyvanpour , An Analytical tool for data stream mining techniques based on challenges and requirements" In : International Journal of Engineering Science and Technology (IJEST) March ,2011

[24] Maghesh Gudipathy , Big data testing Approach to overcome Quality Challenges: In : Infosys Labs Briefing , Volume 11, 2013.

[25] Data Security Challenges Oracle - Security Overview, Part Number A96582-01

[26] Dariusz Brzeziński ,Mining Data Streams With Concept Drift", In : Master's thesis Poznan University of Technology Faculty of Computing Science and Management Institute of Computing Science, 2010.

[27] Data Mining Stream, Time-Series, and Sequence Data

[28] Supun Kamburugamuve ,Survey of Streaming Data Algorithms

[29] Michael Hahsler, Matthew Bolanos, Introduction to stream: An Extensible Framework for Data Stream Clustering Research with R,2017

[30] HangYang ,SimonFong, Incremental Optimization Mechanism for Constructing a Decision Tree in Data Stream Mining , In: Hindawi Publishing Corporation Mathematical Problems in Engineering Volume 2013,ArticleID580397,14 pages ,2013

[31] Mirela Teixeira Cazzolato, Marcela Xavier Ribeiro, Classifying High-Speed Data Streams Using Statistical Decision Trees, In: Journal of Information and Data Management, Vol. 5, No. 1, February 2014, Pages 84–93. 2014

[32] Nicolas Kourtellis Gianmarco De Francisci Morales Albert Bifet Arinto Murdopo, VHT: Vertical Hoeffding Tree, In : arXIV :1607.08325V1 [cs.DC] ,2016

[33] Tomoyasu Takata, Seiichi Ozawa, A Neural Network Model for Learning Data Stream with Multiple Class Labels, In: IEEE Computer Society , 978-0-7695-4607-0/11, 2011

[34] Zeineb Hammami, On line self adaptive framework for tailoring a neural agent learning model addressing dynamic real time scheduling problems, In : The Society of manufacturing Engineers , Elsevier , 0278-6125 ,2017

[35] Yan-Nei Law and Carlo Zaniolo, An Adaptive Nearest Neighbor Classification Algorithm for Data Streams, In : PKDD , LNAI 3721, pp. 108–120, Springer-Verlag Berlin Heidelberg,2005

[36] Jesmin Jahan Tithi, Enabling K-nearest Neighbor Algorithm Using a Heterogeneous Streaming Library: hStreams, In: 16 Nov 13–18, ACM. ISBN 123-4567-24-567/08/06. DOI: 10.475/123,2016

[37] Isah A Lawala,Salihu A.Abdulkarim,, Adaptive SVM for Data Stream Classification, , In: SACJ, ISSN 2313-7835 9,2017

[38] Pranamita Nanda,B.Sandhiya,R.Sandhiya, SVM Classifier Algorithm For Data Stream Mining Using Hive And R, In: International Research Journal of Engineering and Technology (IRJET) e-ISSN: 2395 -0056 ,Volume: 04 Issue: 03 | Mar ,2017

[39] C´esar A. Astudillo, Javier I. Gonzlez1, B. John Oommen, and Anis Yazidi, Concept Drift Detection using Online Histogram-based Bayesian Classifiers, In : Springer ,2016

[40] Leandro L. Minku, and Xin Yao, DDD: A New Ensemble Approach For Dealing With Concept Drift, In : Research in Computational Intelligence and Applications (CERCIA), IEEE 2011

[41] Joao Gama, A Survey on concept drift adaptations In : ACM . Volume 46 Issue 4, April 2014 Article No. 44 2014

[42] Parneeta Sidhu, M. P. S. Bhatia, An online ensembles approach for handling concept drift in data streams: diversified , In : Springer DOI 10.1007/s13042-015-0366-1 ,2015

[43] HangYang ,SimonFong, Incremental Optimization Mechanism for Constructing a Decision Tree in Data Stream Mining , In: Hindawi Publishing

November-December 2019
ISSN: 0193-4120 Page No. 4493 - 4502

Corporation Mathematical Problems in Engineering Volume 2013,ArticleID580397,14 pages ,2013

[44] Gang Liu, Hong-rong Cheng, Zhi-guang Qin, Qiao Liu, Cai-xia Liu , E-CVFDT:An Improving CVFDT Method forConcept Drift Data Stream, In: IEEE 978-1-4799-3051-7/13 ,2013

[45] DONG Zhenjiang , LUO Shengmei , WEN Tao, ZHANG Fayang and LI Lingjuan, Random Forest Based VFDT Algorithm for data Stream , In : December , Vol.15 No. S2,2017

[46] Jesmin Jahan Tithi, Enabling K-nearest Neighbor Algorithm Using a Heterogeneous Streaming Library: hStreams, In: 16 Nov 13–18, ACM. ISBN 123-4567-24-567/08/06. DOI: 10.475/123,2016

[47] Kyong Ha Lee, Hyunsik Choi, Bongki Moon, Parallel data processing with MapReduce: a survey. SIGMOD Record 40(4):11-20, 2012

[48] Afrati, F.N.; Borkar, V.; Carey, M.; Polyzotis, N.; Ullman, J.D. Map-Reduce extensions and recursive queries. In Proceedings of the 14th International Conference on Extending Database Technology, Uppsala, Sweden, 22–24 March; pp. 1–8,2011

[49] Saeed Shahrivari, Beyond Batch Processing: Towards Real-Time and Streaming Big Data", In: Computers(Open Access) 2014, 3, 117-129

[50] Xiufeng Liu , Survey of Real-time Processing Systems for Big Data ,2014

[51] Babak Yadranjiaghdam,A Survey on Real-time Big Data Analytics: Applications and Tools, In : International Conference on Computational Science and Computational Intelligence.

[52] DiegoGarcía-Gil1, "A comparison on scalability for batch big data processing on Apache Spark and Apache Flink", In: Big Data Analytics(Open Access), 2017

[53] Zaharia, M.; Chowdhury, M.; Das, T.; Dave, A.; Ma, J.; McCauley, M.; Franklin, M.J.; Shenker, S.; Stoica, I. Resilient distributed datasets: A fault-tolerant abstraction for in-memory cluster computing. In Proceedings of the 9th USENIX Conference on Networked Systems Design and Implementation, San Jose, CA, USA, 25–27 April ,2012

[54] Saeed Shahrivari ,Beyond Batch Processing: Towards Real-Time and Streaming Big Data Computers 3, 117-129; doi:10.3390/computers3040117,2014

[55] https://moa.cms.waikato.ac.nz

[56] Hassan Nazeer, Real-time Text Analytics Pipeline Using Open-source Big Data Tools

[57] Adnan Akbar, "Predictive Analytics for Complex IoT Data Streams", IEEE Internet Of Things Journal

[58] Balaji,S., Meghana Patil, Carolyn McGregor AM, A Cloud based Big Data Based Online Health Analytics for Rural NICUs and PICUs in India: Opportunities and Challenges, In: IEEE 30th International Symposium on Computer-Based Medical Systems

[59] Thanga prasad .S.,Sangavi. S, Deepa. A, Sairabanu. F, Ragasudha. Diabetic Data Analysis In Big Data With Predictive Method

[60] Basil Shaik, HUMAN: Hadoop Used Medical Analytics : A Survey, In: International Conference On Big Data Analytics and computational Intelligence ,2017

[61] Bendre M.R., Thool .R.C, Big Data in Precision Agriculture : Weather Forecasting for Future Farming In : 1st International Conference on Next Generation Computing Technologies (NGCT-2015) Dehradun, India, 4-5 September ,2015

[62] Ponce-Guevara K.L,GreenFarm-DM: A tool for analyzing vegetable crops data from a greenhouse using data mining techniques In IEEE,2015

[63] Suhas Athani, CH Tejeshwar, Soil moisture monitoring using IoT enabled arduino sensors with neural networks for improving soil management for farmers and predict seasonal rainfall for planning future harvest in North Karnataka – India , In :International conference on I-SMAC (IoT in Social, Mobile, Analytics and Cloud,2017

[64] RabiaLatif, HaiderAbbas, SeemabLatif,1 andAshrafMasood1 , EVFDT: An Enhanced Very Fast Decision Tree Algorithm for Detecting Distributed Denial of Service Attack in Cloud-Assisted Wireless Body Area Network, In: Hindawi Publishing Corporation Mobile Information Systems , Article ID 260594, 13 pages ,2015

[65] Bobin K Sunny, Janardhanan P S, Anu Bonia Francis and Reena Murali, Implementation of a Self-Adaptive Real Time Recommendation System using Spark Machine Learning Libraries, In: IEEE

SPICES 1570358660 ,2017

[66] Kazem Fathi1, Sayyed Majid Mazinani2, Combining KNN and Decision Tree Algorithms to Improve Intrusion Detection System Performance, In: Recent Advances in Communications, ISBN: 978-1-61804-318-4

[67] Kalpesh Adhatrao, Aditya Gaykar, Amiraj Dhawan, Rohit Jha and Vipul Honrao, Predicting Students' Performance Using ID3 And C4.5 Classification Algorithms, International Journal of Data Mining & Knowledge Management Process (IJDKP) Vol.3, No.5, September , 2013

[68 Simon Fong, Yang Hang, Sabah Mohammed and Jinan Fiaidhi, Stream-based Biomedical Classification Algorithms for Analyzing Biosignals, Journal of Information Processing Systems, Vol.7, No.4, December ,2011