THE USE OF SPARK FOR NEAREST NEIGHBOR CLASSIFICATION FOR HIGH-SPEED BIG DATA STREAMING

Asogwa Samuel C[1], Onwuzo J.C[2], Chigbundu Kanu Enyioma[3], Obayi Adaora[4]

[1, 2, 3] Department of Computer Science. Michael Okpara University of Agriculture Umudike. Abia state.

[4]Department of Computer Science, University of Nigeria Nsukka

ABSTRACT

One of the most complicated and difficult aspects of machine learning and Artificial Intelligence (AI) application is data mining and parallel data transmission. This necessitates the adoption of methods that demonstrate high process effectiveness and efficiency, as well as the ability to change their structure and manage a large variety of data that arrives on a regular basis. This should lead to the presentation of a new incremental and distributed classifier based on the preferred closest neighbor algorithmic rule. The research employs the K-NN machine learning classifier, which is well-suited to such a demanding setting. This approach, which is implemented in Apache Spark, uses a distributed metric-space ordering to speed up searches. A great deal of live data transmission containing useful data, referred to as big data, is created frequently. For handling such large volume of data, there is a necessity of big data structures, which the Apache Spark resolves in this work. The model will performs up to one hundred circumstances speedier than ancient systems like Hadoop Map reduce and the system will be used for online payment processing and big data search pattern.

Keyword: *Personal Computer (PC), Classification, Big data, Machine learning (ML). Apache Spark*

1. INTRODUCTION

The monstrous volume of data assembled by contemporary systems became inescapable, as many examination exercises require gathering progressively tremendous measures of information. For example, Large Hadron Collider experiments produces 30 petabytes of data per year. Likely applications for enormous information investigation procedures could be found in every human action space. Endeavors might want to find fascinating customer behavior attributes, e.g., on the basis of sensor or Internet information. Another model would be works on personalized clinical therapy for singular patients based on his/her clinical records, such as clinical history, genomic, cell, and natural information. The contemporary man is encircled by huge volumes of information showing up consistently from various sources, accordingly one may say that we are living in the big information period. Big information is normally portrayed by the purported 5V's (Volume, Velocity, Variety, Veracity, and Value), which describe its enormous volume, dynamic nature, various structures, various characteristics and value for human [1].

As a rule we don't manage static information assortments, but rather with dynamic ones. They show up in a type of nonstop batches of information, called information streams [2]. In such situations we need not exclusively to deal with the volume, but likewise the speed of information, continually refreshing our learning model and adjusting it to the present status of the stream. To add a further trouble, numerous cutting edge information sources create their yields with extremely short spans, hence prompting high-speed information streams [3]. In this work we will essentially zero in on two attribute of the big information, i.e., volume and speed. Huge information must be investigated proficiently and changed over into valuable information which could be utilized by undertakings (among others) to build their upper hand [4]. Nonetheless, there exist a considerable hole between contemporary

handling and storage limits that show our ability to catch and store information has far outperformed our ability to measure and use it.

Moore's law says that preparing limit double at regular intervals, while plate storage limit doubles like clockwork (storage law) [5]. It could cause that supposed information tombs, i.e., volume of information which are stored but never dissected, may show up. Along these lines, we need to foster devoted tools and methods which can examine colossal volume of approaching information and moreover to contemplate that each record might be broke down just a single time to decrease the general computing costs [6]. MapReduce was the principal programming worldview in managing the wonder of big information, presented in 2003 [7]. As of late, another huge scope handling system, called Apache Spark [8], [9], is acquiring significance in the big information area because of its great performance in iterative and steady techniques. Languid learning [10] (likewise called occasion based learning) is considered as one of the least difficult and best plans in supervised learning [11], in which generalization is conceded until an inquiry is put forth to the defense base. Be that as it may, as distance between each pair of cases must be figured (quadratic intricacy), these techniques will in general have much slower grouping.

The problems of information produced by system without being properly put into utilization is really influencing the whole global system. The regular operating system has cause the abuse of the systems usage, realizing without a doubt that a similar information are produced within a given period of time, can cause the overabundance use of a system. The majority of the yield of the system can be anticipated. Coming up next are the difficulties of the current system; the current need forecast of result and continually extreme use system to produce same information, Lack association in information age, need distribution in its system, just accessible in privately based systems, need speed in search measure thereby causing languid heap of information and so forth

The limit real boils down to information transmission and limit of information to be dissect. The big information and information mining of huge and high speed transmission. In this proposal, the system is simply limited to a big information examination.

The importance is identified with creating an information from given system and dissect it for future expectation of a connected information system. To this information are not, at this point consistently created but examined and utilized for future expectation. Expectation the possible next occurrence of search record in a pool of high speed information streaming.

2.1 Review Overview

At the point when it comes the central passed on systems for titanic scale data managing the so to speak title come educate which is Google, which is careful for laid out the Layout Reduce in 2003[12]. Design Lessen associations the bunches of PCs are utilized for subsequently organizing data. Mappers and Reducers are the two boundaries that customer should finish in Layout Diminish. In Diagram Organize the key – respect sets are thought about unmistakably from dispersed record structure. These are change into another game plan of sets. Center are looking at and changing a ton of sets from in any event one data zones. In Diminish Stage customer depicted cutoff points are utilized to send the key correspondent sets and established to engage the incredible surrender. For more information border Diagram Diminish and others dispersed designs, on the off chance that it's not all that much burden check [13]. Another remarkably outstanding open source utilization of Layout Reduce is Apache Hadoop [14] [15]. Its tried and true, adaptable and spread figuring. Hadoop is having a limitations that is isn't well fitting for where there's need for express data reuse. For blueprint online instinctually, just as iterative handling are influenced by this issue [16].

2.1.1 Machine Learning Approach for Data Mining Streaming

Man-made intelligence figuring are working on tremendous volume of data. Speed of the data time likewise plays the fundamental fragment where AI tally are enormous. With the headway in material, progressed breaking point, and web and correspondence propels, plan ML ask generally and progress which outperform needs in appear, check and hypothesis sorts of progress are in the long run tried by the making transcendence of monstrous data aggregations, for instance, a few hours video-sharing area each second, or petabytes of online life on at least billion customer social constructions. The climb of huge data to boot being went with by expanding throbbing for higher dimensional and progressively complex ML models with billions to trillions of boundaries. In figure out to back the reliably expanding multi-layered nature of data, or to encourage still higher perceptive precision (for example for better customer benefit and mending end) and bolster even more shrewdly undertakings (for example driver less vehicles and semantic diagram of video data). Organizing such enormous ML models over such gigantic data is past the breaking point and computation abilities of single machine. This opening has influenced a making assemblage of sometime later work in scattered ML, where ML projects are executed over take a gander at gatherings, worker homesteads, and cloud providers sagacity tens to thousands of machine.

2.1.2 Data Stream Mining

Events may come continually in a structure of a possibly boundless data stream [17]. This makes unused tending to for learning calculations, as they should offer change rebellious for clearing enlightening file [18]. Moved obstacles on an exceptionally basic level be taken into contemplating that are not show up or not too major in torpid circumstances [19]. Understudy must have moo response and redesign times, as forefront things must be managed with as some time actually long as they wrapped up accessible. Likewise long organizing would cause a deferral, as stacking showing up things would considering the way that it were addition with the stream progress [20]. Furthermore, spilling checks must perceive obliged limit space and memory necessities. One can't store all of articles from a stream, as data volume will perpetually make [21]. Along these lines, objects found the opportunity to be coordinated of in the wake of managing and understudy must not need a get to beginning at now observed events. Data streams are regularly described which is thought of and called thought drift [22], [23]. It might be described as a distinction in brand name in drawing closer data all through stream managing. In spilling conditions [24], the drawing nearer articles show up progressively, subsequently data streams can be arranged in two shifting movement modes.

1) Chunk (collect), where data touch base in a condition of event squares or we accumulate adequate events to shape one.

2) Online, where events show up separately and we should design them as some time actually long as they ended up open.

There are some possible approaches to manage acquiring from data streams.

1) Adjusting the classifier whenever unused data finds the opportunity to be available.

2) Utilizing a sliding window approach.

3) Utilizing a continuous or online understudy.

The key of explored strategies is far away from being fitting in an honest to goodness stream mining condition. Orchestrating a bleeding edge portray at anything point a cutting edge set of events show up would propel prohibitive computational costs and over the best need for a cutoff space in organize to oblige the consistently creating level of the masterminding set. Likewise, in the midst of the orchestrating handle the classifier would be shut off for data overseeing, which would incite a key time delay. These parts drive us to design specific systems that don't progress forward from the said obstacles. Sliding window-based classifiers were engineered from a general perspective for floating data streams, as they partner the overlooking part in figure out to organize of immaterial tests and acclimate to showing up changes [25].

2.1.3 K-Nearest Neighbor Algorithm

Nearest Neighbor Algorithm is a straightforward algorithm utilized in information mining. The job of this algorithm is to peruse the information and sort them with respect to given condition. Accordingly when a non-existed input information is given it checks the number of nearest distance record 'K' to the information gave. With respect to the

nearest information to the info, the yield is figured. This algorithm is suitable just for little informational collections as it gives boisterous information when it is carried out in an immense informational collection. The K worth is given by the client which is the number of test existing record that ought to be contemplated [30]. The below shown graphs are straightforward portrayals of K-NN algorithm. The diagram's x-pivot means the temperature of the human body and y-hub indicates natural eye tone. At first the current information are plotted in the chart as Disease A and Disease B by thinking about the illness attributes. At the point when an obscure infection x is to be resolved, the algorithm contrasts the current plot and respect to nearest separate records and predicts a yield.

2.1.4 Conceptual Review

The reviewed concept, is a prediction of process using a machine learning algorithm K-NN for prediction. This involve a large high speed streaming data for a classification of data mined. The system is a web based platform designed and implemented using python flask framework in connection with analytical libraries like pandas, numpy, sckit-learning etc. The spark method is applied in this platform to actualize the system objectives.

2.2 Empirical Review

Apache Start probably could be a speedy and broadly valuable bunch figuring system. It gives abnormal state APIs in Java, Scala, Python and R, and a smoothed out motor that fortifies basic execution diagrams. It too braces an affluent plan of higher-level devices including Spark SQL for SQL and figured out data preparing, MLlib for AI, GraphX for chart managing, and Spark Spouting. Flexible Scattered Datasets (RDD) may be a basic data design of Start. It is an invariable scattered collection of articles. Each dataset in RDD is bound into strong tasks, which may be figured on different focal points of the gathering. RDDs can contain any sort of Python, Java, or Scala objects, tallying customer portrayed classes. RDD can be a perused just, isolated assembling of records. RDDs can be made through deterministic assignments on one or the other information on reliable breaking point or diverse RDDs. RDD may be an inadequacy tolerant social event of fragments that can be managed in equal. There are two distinct approaches to shape RDDs parallelizing a current amassing in your driver program, or referring to a dataset in an external cutoff system, for instance, a common record structure, HDFS, HBase, or any data source propelling a Hadoop Input Orchestrate. Begin utilizes the possibility of RDD to recognize speedier and persuading MapReduce assignments.

The vital thought of Start is Resilient Distributed Datasets (RDD); it bolsters in-memory managing estimation. This infers, it stores the state of memory as a test over the occupations and the question is sharable between those jobs. Data participating in memory is 10 to various occasions faster than mastermind and Disk. Begin what's more honors us to the RDD's API in spilling condition through the difference in data streams into little gatherings. Begin Streaming's organize enables a comparative group code to be utilized in spouting assessment, without a fundamental for principal changes. The AI library of Start is having sevral bundles in MLlib which consolidation learning figurings and utilities [26] [27]. Course of action, headway, apostatize, cooperative filtering, bunching and data pre-handling are the specific assignments should be possible on this Mllib

In reality, this segment audits the movement of works done the writing explored on the examination postulation. What different analysts have done around there, the hole those specialists left in their work. Accordingly, the hole your work is filling. Put the alternate path round, your contribution to information.

2.3 Knowledge Gap

The existing system perform search, but the proposed system is intended to use a machine search K-NN model to improve of search process. A faster means to identify and locate a search index

3.1. METHODOLGY

This research work is taking a dimension of quantitative research method which involve the use of dataset collected from a very large high speed streaming data.

3.2 Research Model (Spiral Model)

Research can be defined as "an activity that involves finding out, in a more or less systematic way, things you did not know" Walliman and Walliman, (2011). "Methodology is the philosophical framework within which the research is conducted or the foundation upon which the research is based" Brown, (2006). Research Methodology chapter of a research describes research methods, approaches and designs in detail highlighting those used throughout the study, justifying my choice through describing advantages and disadvantages of each approach and

design taking into account their practical applicability to our research. O'Leary (2004) describes methodology as the framework which is associated with a particular set of paradigmatic assumptions that we will use to conduct our research. Allan and Randy (2005) insist that when conducting a research methodology should meet the following two criteria:

The spiral model combines the idea of iterative development with the systematic, controlled aspects of the waterfall model. This Spiral model is a combination of iterative development process model and sequential linear development model i.e. the waterfall model with a very high emphasis on risk analysis. It allows incremental releases of the product or incremental refinement through each iteration around the spiral.

3.2.1 Spiral Model - Design

The spiral model has four phases. A software project repeatedly passes through these phases in iterations called Spirals.

- **Identification**

This phase starts with gathering the business requirements in the baseline spiral. In the subsequent spirals as the product matures, identification of system requirements, subsystem requirements and unit requirements are all done in this phase. This phase also includes understanding the system requirements by continuous communication between the customer and the system analyst. At the end of the spiral, the product is deployed in the identified market.

- **Design**

The Design phase starts with the conceptual design in the baseline spiral and involves architectural design, logical design of modules, physical product design and the final design in the subsequent spirals.

- **Construct or Build**

The Construct phase refers to production of the actual software product at every spiral. In the baseline spiral, when the product is just thought of and the design is being developed a POC (Proof of Concept) is developed in this phase to get customer feedback. Then in the subsequent spirals with higher clarity on requirements and design details a working model of the software called build is produced with a version number. These builds are sent to the customer for feedback.

- **Evaluation and Risk Analysis**

Risk Analysis includes identifying, estimating and monitoring the technical feasibility and management risks, such as schedule slippage and cost overrun. After testing the build, at the end of first iteration, the customer evaluates the software and provides feedback. The following illustration is a representation of the Spiral Model, listing the activities in each phase.

Based on the customer evaluation, the software development process enters the next iteration and subsequently follows the linear approach to implement the feedback suggested by the customer. The process of iterations along the spiral continues throughout the life of the software.

3.2.2 Spiral Model Application

The Spiral Model is widely used in the software industry as it is in sync with the natural development process of any product, i.e. learning with maturity which involves minimum risk for the customer as well as the development firms. The following pointers explain the typical uses of a Spiral Model −

- When there is a budget constraint and risk evaluation is important.
- For medium to high-risk projects.
- Long-term project commitment because of potential changes to economic priorities as the requirements change with time.
- Customer is not sure of their requirements which is usually the case.
- Requirements are complex and need evaluation to get clarity.
- New product line which should be released in phases to get enough customer feedback.
- Significant changes are expected in the product during the development cycle.

3.2.3 Spiral Model - Pros and Cons

The advantage of spiral lifecycle model is that it allows elements of the product to be added in, when they become available or known. This assures that there is no conflict with previous requirements and design. This method is consistent with approaches that have multiple software builds and releases which allows making an orderly

transition to a maintenance activity. Another positive aspect of this method is that the spiral model forces an early user involvement in the system development effort.

On the other side, it takes a very strict management to complete such products and there is a risk of running the spiral in an indefinite loop. So, the discipline of change and the extent of taking change requests is very important to develop and deploy the product successfully.

The advantages of the Spiral SDLC Model are as follows −

- Changing requirements can be accommodated.
- Allows extensive use of prototypes.
- Requirements can be captured more accurately.
- Users see the system early.
- Development can be divided into smaller parts and the risky parts can be developed earlier which helps in better risk management.

The disadvantages of the Spiral SDLC Model are as follows −

- Management is more complex.
- End of the project may not be known early.
- Not suitable for small or low risk projects and could be expensive for small projects.
- Process is complex
- Spiral may go on indefinitely.
- Large number of intermediate stages requires excessive documentation.

3.3 Proposed Model of the New System

The proposed model is an architectural design of the overview of system flow and operational steps.

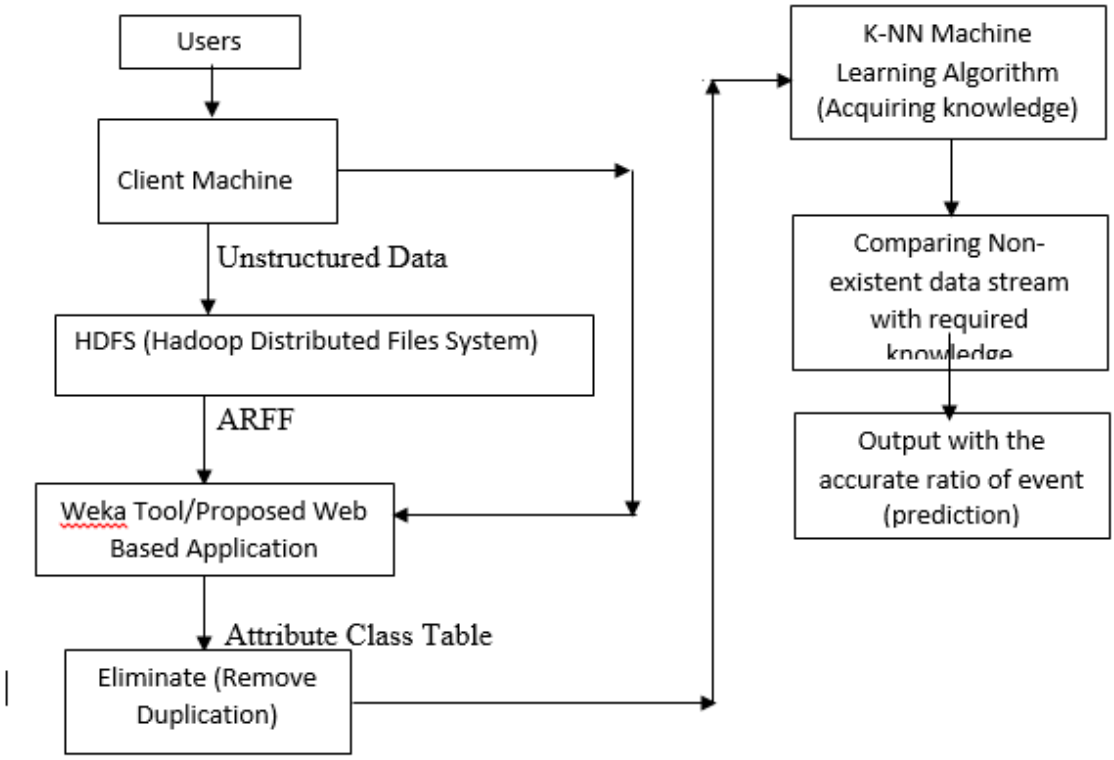

Figure 3.0 Overview of the proposed system

The proposed system undergoes through seven important steps which are shown in the above data flow diagram are listed below:-

Step 1: The user uploads the huge amount of unstructured and semi-structured data into the client machine.
Step 2: The client machine stores the received data in the HDFS (Hadoop Distribute File System).
Step 3: The format of the unstructured data is converted into to Comma-Separated Systems format followed by converting them again into ARFF file format.
Step 4: Using the service of the HDFS, the Weka tool/Web based system mines each attribute and creates an attribute relation table using Job Tracker and Task Node services.
Step 5: Now a Reduction technique is applied to eliminate the duplication in the data streams.
Step 6: Naive Bayes algorithm is implemented on the reduced attribute relation table by which the algorithm acquires knowledge from the data streams in the form of probabilities.
Step 7: Now a non-existing input data streams is taken and compared with the acquired knowledge to predict its output.
Step 8: An output is generated with the accurate ratio of the event outcome

4.0 RESULTS AND DISCUSSIONS

We have demonstrated a nearest neighbor order game-plan for arranging huge and tall speed information streams using Apache Start. Commonly gainful methodology for expansive scale, tall speed and gushing issues. This proposed framework composes the occasions by making a scattered assignment reliance diagram by using metric tree. In this tree involving beat level tree courses the inquiries to the leaf focuses and a lot of dispersed sub outlines that plays out the see parallel. This proposed framework deals with choosing the occasion that produces walks the execution. As future work we are going make update in this work by in light of the fact that it were permitting the development of adjust bits of information and expelling the outdated ones. This will make progress the abundance of student.

4.1 Prediction process

Classification process is a rough capacity that is begun when new unlabeled information show up at the system. For every component the algorithm looks for the nearest leaf hub in the expert hub and mixes the components to the slave machines. Then, the standard M-tree search process is utilized to recover the kp-neighbors of each new component. For each gathering, framed by another component and its neighbors, the algorithm predicts the component's class by applying the larger part casting a ballot plan to its neighbors.

5.0. RECOMMENDATIONS

The proposed system is just a basic initiation taken to analyze and study data as analysis of huge data streams is at the beginning stage for the current generation. The future scope of this paper involves implementing many different statistical and relative algorithm on the data streams to increase the better study and analysis of the data which intern enhance the accuracy of the computer to acquire knowledge deeply and predict the outcome of hypothesis events. Many several machine learning algorithm are very much available, the adoption of other model are as well welcome as the future research advancement of the speedy and faster means of identifying a search index.

Reference

[1] V. Mayer-Schnberger and K. Cukier, Big Data: A Revolution That Will Transform How We Live, Work and Think. Viktor Mayer-Schnberger. John
Murray Publishers, 2013.
[2] J. Gama, Knowledge Discovery from Data Streams. Chapman & Hall/CRC, 2010.
[3] D. Han, C. G. Giraud-Carrier, and S. Li, “Efficient mining of high-speed uncertain data streams,” Applied Intelligence, vol. 43, no. 4, pp. 773–785, 2015.
[4] X. Wu, X. Zhu, G. Wu, and W. Ding, “Data mining with big data,” IEEE Transactions on Knowledge and Data Engineering, vol. 26, no. 1, pp. 97–107,
2014.

[5] U. Fayyad and R. Uthurusamy, "Evolving data into mining solutions for insights," Communications of ACM, vol. 45, no. 8, pp. 28–31, Aug. 2002.
[Online]. Available: http://doi.acm.org/10.1145/545151.545174
[6] A. Fern´andez, S. del R´ıo, V. L´opez, A. Bawakid, M. J. del Jes´us, J. M. Ben´ıtez, and F. Herrera, "Big data with cloud computing: an insight on the
Co mputing environment, mapreduce, and programming frameworks," Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery, vol. 4,
no. 5, pp. 380–409, 2014.
15
[7] J. Dean and S. Ghemawat, "Mapreduce: Simplified data processing on large clusters," in OSDI 2004, 2004, pp. 137–150.
[8] H. Karau, A. Konwinski, P. Wendell, and M. Zaharia, Learning Spark: Lightning-Fast Big Data Analytics. O'Reilly Media, Incorporated, 2015.
[9] Apache Spark: Lightning-fast cluster computing, "Apache Spark," 2017, [Online; accessed January 2017]. [Online]. Available: https://spark.apache.org/
[10] D. Aha, Lazy Learning. Kluwer Academic Publishers, 1997.
[11] C. C. Aggarwal, Data Mining: The Textbook. Springer, 2015.

[12]. J. Dean and S. Ghemawat, "Mapreduce: Simplified data processing on large clusters," in Proc. OSDI, San Francisco, CA, USA, 2004, pp. 137–150.

[13]. A. Fernández et al., "Big data with cloud computing: An insight on the computing environment, mapreduce, and programming frameworks," Wiley Interdiscipl. Rev. Data Min. Knowl. Disc., vol. 4, no. 5, pp. 380–409, 2014.

[14]. T. White, Hadoop, the Definitive Guide. Sebastopol, CA, USA: O'Reilly Media, 2012.

[15]. Apache Hadoop Project. (2017). Apache Hadoop. [Online]. Accessed on Jan. 2017. [Online]. Available: http://hadoop.apache.org/

[16]. J. Lin, "Mapreduce is good enough? If all you have is a hammer, throw away everything that's not a nail!" Big Data, vol. 1, no. 1, pp.

28–37, 2012.

[17] X. Meng et al., "Mllib: Machine learning in apache spark," J. Mach. Learn. Res., vol. 17, no. 1, pp. 1235–1241, 2016.

[18]. A. Spark. Machine Learning Library (MLlib) for Spark. Accessed on Jan. 2017. [Online]. Available: http://spark.apache.org/docs/latest/mllib-guide.html.

[19]. T. M. Cover and P. E. Hart, "Nearest neighbor pattern classification," IEEE Trans. Inf. Theory, vol. 13, no. 1, pp. 21–27, Jan. 1967.

[20]. H. Samet, Foundations of Multidimensional and Metric Data Structures (The Morgan Kaufmann Series in Computer Graphics and Geometric Modeling). San Francisco, CA, USA: Morgan Kaufmann, 2005.

[21]. A. Gionis, P. Indyk, and R. Motwani, "Similarity search in high dimensions via hashing," in Proc. 25th Int. Conf. Very Large Data Bases (VLDB), Edinburgh, U.K., 1999, pp. 518–529.

[22]. M. M. Gaber, "Advances in data stream mining," Wiley Interdiscipl. Rev. Data Min. Knowl. Disc., vol. 2, no. 1, pp. 79–85, 2012.

[23]. B. Krawczyk, L. L. Minku, J. Gama, J. Stefanowski, and M. Woźniak, "Ensemble learning for data stream analysis: A survey," Inf. Fusion, vol. 37, pp. 132–156, Sep. 2017.

[24]. A. Bifet, G. D. F. Morales, J. Read, G. Holmes, and B. Pfahringer, "Efficient online evaluation of big data stream classifiers," in Proc. 21Th ACM SIGKDD Int. Conf. Knowl. Disc. Data Min., Sydney, NSW, Australia, 2015, pp. 59–68.

[25]. S. Ramírez-Gallego, B. Krawczyk, S. García, M. Woźniak, and F. Herrera, "A survey on data preprocessing for data stream mining: Current status and future directions," Neurocomputing, vol. 239, pp. 39–57, May 2017.

[26]. J. Gama, I. Žliobaite, A. Bifet, M. Pechenizkiy, and A. Bouchachia, "A survey on concept drift adaptation," ACM Comput. Surveys, vol. 46, no. 4, pp. 1–37, 2014.

[27]. C. Alippi, D. Liu, D. Zhao, and L. Bu, "Detecting and reacting to changes in sensing units: The active classifier case," IEEE Trans. Syst., Man, Cybern., Syst., vol. 44, no. 3, pp. 353–362, Mar. 2014.

[28]. Z. Pervaiz, A. Ghafoor, and W. G. Aref, "Precision-bounded access control using sliding-window query views for privacy-preserving data streams," IEEE Trans. Knowl. Data Eng., vol. [29], no. 7, pp. 1992–2004, Jul. 2015.

[30] J. Maillo, S. Ramirez, I. Triguero, and F. Herrera, NN-IS: An iterative spark-based design of the k-nearest neighbors classifier for big data, Knowledge. Based Syst., vol. 117, pp. 15, Feb. 2017.

Allan, AJ, Randy, LJ, 2005, Writing the Winning Thesis or Dissertation. A Step-by-Step Guide, Corwin Press, California

Brown RB, 2006, Doing Your Dissertation in Business and Management: The Reality of Research and Writing, Sage Publications

Cohen, L, Manion, L, Morrison, K & Morrison, RB, 2007, Research Methods in Education, Routledge

O'Leary Z. 2004 " The essential guide to doing research". Sage.

Walliman, N. S. & Walliman N. (2011) "Research methods: the basics" Taylor and Francis