Journal of Autonomous Intelligence (2019) Volume 2 Issue 2.
doi:10.32629/jai.v2i3.59

Original Article

An Experimental Analysis of the Applications of Datamining Methods on Bigdata

CH. Naga Santhosh Kumar[1*], K.S. Reddy[1]

[1]Professor CSE, ANURAG Engineering College, JNUTH, KODADA, India; Email: santhosh.goal19@gmail.com

[2]Researcher, Hyderabad India; Email: sudheercse@gmail.com

ABSTRACT

Data mining is a procedure of separating covered up, obscure, however possibly valuable data from gigantic data. Huge Data impactsly affects logical disclosures and worth creation. Data mining (DM) with Big Data has been broadly utilized in the lifecycle of electronic items that range from the structure and generation stages to the administration organize. A far reaching examination of DM with Big Data and a survey of its application in the phases of its lifecycle won't just profit scientists to create solid research. As of late huge data have turned into a trendy expression, which constrained the analysts to extend the current data mining methods to adapt to the advanced idea of data and to grow new scientific procedures. In this paper, we build up an exact assessment technique dependent on the standard of Design of Experiment. We apply this technique to assess data mining instruments and AI calculations towards structure huge data examination for media transmission checking data. Two contextual investigations are directed to give bits of knowledge of relations between the necessities of data examination and the decision of an instrument or calculation with regards to data investigation work processes.

Keywords: *Data Mining; Big Data; Knowledge Discovery Databases; Decision Tree; Cloud Data Mining; K-Closest Neighbor; Artificial Intelligence; Cluster*

ARTICLE INFO
Received: Oct 26, 2019
Accepted: Dec 23, 2019
Available online: Dec 24, 2019

***CORRESPONDING AUTHOR**
CH. Naga Santhosh Kumar, Professor CSE, ANURAG Engineering College ,JNUTH , KODADA, INDIA;
santhosh.goal19@gmail.com;

CITATION
CH. Naga Santhosh Kumar and K.S. Reddy. An Experimental Analysis Of The Applications Of Datamiing Methods On Bigdata. Journal of Autonomous Intelligence 2019; 2(2): 64-72. doi: 10.32629/jai.v2i3.59

1. Introduction

Enormous data is the new reality in the telecom world. Over late years, the portable broadband traffic has had a hazardous development because of across the board appropriation, progressed new systems, expanding entrance of cell phones, and a great many versatile applications. This development will proceed at a fast pace as expanding organization of Internet of Things, sharable, uploadable and findable substance by portable clients, sensors, associated autos, etc. Portable enormous data has demonstrated valuable for limit and execution observing (e.g., during typical activity or under huge occasions), investigating, reasonable lab testing, recreation, new component structure, a building advancement of versatile system foundation items.

Data digging is a strategy for finding intriguing examples just as clear and reasonable models from enormous scale data. Data mining can be utilized to discover connections or examples among many fields in huge social database[1]. Data mining is likewise the way toward finding or discovering some new, substantial, justifiable, and possibly valuable types of data. Cloud data mining (CDM) is a repetitive procedure that requires an exceptional framework dependent on use of new stockpiling innovations, taking care of, and handling. Huge Data/Hadoop is the most recent promotion in the field of data preparing. Through the combination of inside and out examination of (data mining) and distributed computing, arrangements getting to data mining

administrations without fail and all over and from different stages and gadgets will be made conceivable[2].

Data mining is the way toward breaking down data collections in novel approaches to discover unsuspected connections. The connections also, outlines determined through data mining are frequently alluded to as models or examples that concentrate understood, obscure and potential helpful data from data. This is required so as to anticipate future patterns and practices, to settle on proactive choices, and to respond to business addresses that devour an excessive amount of time to reply. Various data mining systems have been concentrated to process and to dissect a few kinds of data designs, where the most well known data mining undertakings are arrangement, outline, affiliation guidelines mining, and grouping.

Huge data term is a quick developing documentation alluding to the accumulations of the colossal data indexes that can't be prepared utilizing customary database the board frameworks and existing strategies. Huge data present new methodologies for data stockpiling, preparing models, investigation and representation of such gigantic data size inside an acknowledged time term that can be accomplished with average computational frameworks. This is expected to the chiefly described 4Vs; 1) Volume, which demonstrates managing immense measure of data regarding petabytes scale accumulations. 2) Variety, where the classification of huge data has a place with organized, semi-organized, or unstructured data. 3) Velocity, which alludes to the speed of data age or how quick the data are required for handling to fulfill the need. 4) Veracity, which alludes to the irregularity and the low nature of data that can be identified in monstrous data indexes, influencing the preparing of data.

Since the web of things and propelled data advances (for instance, radio recurrence distinguishing proof (RFID) labels and shrewd sensors) are broadly utilized in assembling endeavors for their day by day generation and the board, product lifecycle management (PLM) forms produce a gigantic measure of data[1]. Moreover, the gathering of verifiable data in enterprise resource planning (ERP), store network the executives (SCM), client relationship the board (CRM), and request the executives framework (OMS), just as the convenient gathered data by the broadly utilized manufacturing execution system (MES) and distributed control system (DCS) added to the sharp increment of data throughout the decades. The time of mechanical Big Data has come.

Pioneers of assembling ventures are winding up progressively keen on profiting their organizations by viably utilizing Big Data[1]. Huge data related advancements, for example, knowledge discovery in databases (KDD) and data mining (DM) have been broadly utilized to upgrade the insight and proficiency of the plan, creation, and administration forms in many assembling scenes, for example, item structure improvement, fabricating process streamlining, generation the executives and enhancement (PMO), generation procedure checking and control, quality administration, CRM, SCM, etc. Intel utilizes Big Data for prescient support of gear and enormously lessens the superfluous hardware stop and inactive time.

2. Literature Review

Choice of appropriate programming applications to complete explicit undertakings has turned out to be trying, because of the fast advancement and accessibility of programming. Various execution assessment systems or techniques have been created to help this choice procedure. Execution assessment is a strategy to decide the qualities and shortcomings of the basic engineering or configuration design. Some assessment strategies created are talked about in (Shanmugapriya, 2012), where they are classified into an early and late assessment. The early assessment strategies are programming assessment techniques that can evaluate the product application dependent on its particular and depiction. They are utilized to break down programming quality properties, for example, dependability, execution, versatility and accessibility. A large portion of these assessment strategies are situation based. Situation based assessment techniques distinguish situations in close communication with the partners and efficiently explore the product engineering dependent on these situations. A portion of the models are Software Architecture Analysis Method (SAAM), Architecture Trade-off Analysis Method (ATAM), Family Architecture Analysis Method (FAAM), and Domain-Specific Software Architecture Comparison Model (DoSAM) (Ionita, 2002) (Kazman R.

a., 1994), (Kazman R. a., 1998), (Bergner, 2005) .

SAAM was first presented in 1993 as a situation based early assessment technique (Dobrica, 2002). The fundamental bit of leeway is this assessment strategy is its versatile structure (Kazman R. a., 1993) that permits change of SAAM's basis configuration to create assessment techniques for explicit prerequisites. The strategy incorporates the six stages or exercises of 1) situation advancement, 2) System Architecture (SA) portrayal, 3) Scenario characterization and prioritization, 4) singular situation assessment, 5) situation cooperation, and 6) in general assessment (Babar, 2004). ATAM is another assessment technique for surveying quality traits, for example, modifiability, movability, extensibility, and integrality. DoSAM is another situation based assessment strategy intended to evaluate programming quality properties like execution, versatility, and accessibility (Bergner, 2005). The late assessment strategies are utilized where the product application is inclined to changes.

An approach is presented in (Tvedt, 2002) to assess programming applications that experience alteration during the execution procedure. It evades framework degeneration by effectively and methodicallly distinguishing and revising deviations from the arranged plan at the earliest opportunity, in view of unequivocal and verifiable building rules. It has the accompanying seven stages as pursues: 1) Select the point of view for assessment; 2) Define and select rules, and set up measurements to be utilized in the assessment; 3) Analyze the arranged structural structure so as to characterize building plan objectives; 4) Analyze the source code so as to figure out the real engineering plan; 5) Compare the genuine structure to the arranged plan so as to distinguish deviations; 6) Formulate change proposals so as to adjust real and arranged structures; and 7) Verify that the structure objective infringement have been revised by rehashing stages 4 through 6.

Other late assessment strategies are talked about in (Lindvall, 2003), (Fiutem, 1998), (Murphy, 1995), (Sefika, 1996). The greater part of these assessment strategies are expected for the assessment of a solitary engineering at a specific point in time. For a situation of looking at instruments, they essentially centered around contrasting their outcomes against a specific investigation work, for instance, the precision of order (Bernardino, 2013), (Al-Shawakfa, 2011). Propelled by the property "Speed" of Big Data investigation, we are increasingly worried about the exhibition of various apparatuses. All the more significantly, as far as anyone is concerned, there comes up short on an orderly investigation of assessing data mining apparatuses that are driven by necessities got from an data examination setting. Along these lines, our investigation can likewise be seen as an encounter report on assessing data mining instruments by following a moderately thorough procedure and applying standards of Design of Experiment (DOE) methods. DOE developed generally for agribusiness, synthetic, and procedure ventures. Thinking about its normal association with trial exercises, appropriate DOE procedures have likewise been utilized in exploratory software engineering. With regards to the product designing field, the primary enthusiasm of applying DOE is by all accounts in programming testing from the designer's viewpoint (Iannino, 1997), (Reilly, 2002), (Zevallos, 1998). Our investigation basically stretches out the appropriateness of DOE to programming correlation from the end client's viewpoint.

3. Methods of Data Mining and Big Data

3.1 Concepts of Data Mining and Big Data

Data mining is a lot of strategies for separating profitable data (designs) from data. It incorporates grouping examination, arrangement, relapse, and affiliation principle learning, and so on. For instance, group examination is utilized to separate articles with specific highlights and gap them into certain classes (bunches) as indicated by these highlights. It is a solo examination technique without preparing data. Grouping can be considered the most significant unaided learning issue. Grouping comprises of looking at the highlights of a recently displayed item and allotting to it a predefined class. A few noteworthy sorts of grouping calculations in data mining are choice tree, k-closest neighbor (KNN) classifier, Naive Bayes, Apriori and AdaBoost[1]. Relapse examination recognizes reliance connections among factors covered up by arbitrariness.

KNN classifiers are a sort of nonparametric strategy

for ordering data items dependent on their k nearest preparing data questions in the data space. The KNN classifiers don't develop any classifier model unequivocally; rather they keep all preparation data in memory. Consequently they are not managable to huge data applications.

Data mining administrations abuse and are based over a cloud foundation and other most conspicuous huge data handling advancements to offer functionalities, for example, elite full content hunt, data ordering, arrangement and grouping, coordinated data sifting and combination, and important data accumulation. Propelled content mining strategies, for example, named substance acknowledgment, connection extraction, and conclusion mining help separate important semantic data from unstructured writings. Keen data mining strategies that are being utilized incorporate nearby example mining, similitude learning, and chart mining.

In spilling data mining, Very Fast Decision Tree (VFDT) is a gushing data classifier which begins with just the root hub, sorts preparing data to leaf hubs, and parts the leaf hubs that meet the parting criteria on-the-fly. It very well may be effectively connected to stream data, however it has a few confinements to apply huge data on the grounds that the quality estimates like the data gain for parting traits are assessed over (yet enormous) data subsets.

A method for accelerating the mining of gushing students is to appropriate the preparation procedure onto a few machines. Hadoop is such a programming model and programming structure. Apache S4 is a stage for preparing persistent data streams. S4 applications are intended for consolidating streams and preparing components progressively.

In huge data mining and examination, a few devices and well known open source activities are as per the following:

1) Apache Mahout: Scalable AI and data mining programming dependent on Hadoop. It has executions of grouping, order, collective sifting, and continuous example mining.
2) MOA: Stream data mining programming to perform data mining progressively. It has usage of grouping, order, relapse, visit thing set mining, and successive chart mining.
3) R: open source programming language and programming condition intended for factual processing, data mining/investigation, and representation.
4) GraphLab: abnormal state chart parallel framework worked without utilizing MapReduce.
5) Excel: It gives ground-breaking data handling and factual investigation abilities.
6) Rapid-I Rapidminer: Rapidminer is open source programming utilized for data mining, AI, and prescient examination. Data mining and AI projects given by RapidMiner incorporate Extract, Transform, and Load (ETL); data pre-preparing and representation; demonstrating, assessment, and arrangement.

4. Data Mining for Big Data

Data mining is the way toward discovering connections or examples among many fields in huge social database. Data mining (here and there called data or learning disclosure) is the way toward investigating data from alternate points of view and abridging it into helpful data. In fact, Data mining as a term utilized for the particular classes of six exercises or assignments as pursues: 1) Classification 2) Estimation 3) Prediction 4) Association rules 5) Clustering 6) Description 1) Classification[5] is a lot of methods which are planned for perceiving classifications with new data focuses. As opposed to bunch examination, a characterization strategy uses preparing data indexes to find prescient connections. 2) Estimation manages persistently esteemed results. Given some data, we go through estimation to accompany an incentive for some obscure constant factors, for example, pay, stature or charge card balance. 3) Prediction It's an announcement about the manner in which things will occur later on, regularly yet not constantly founded on involvement or learning. Expectation might be an announcement wherein some result is normal. 4) Association Rules Association standard learning[6,7] is set of methods intended to recognize significant connections or affiliation rules among factors in databases. 5) Clustering Cluster investigation[8] depends on standards of similitudes to characterize objects. This strategy has a place with

unaided realizing where preparing data is utilized.

4.1 Features of Big Data

Huge data comprises of number of highlights. They are:

It is gigantic in size.

Its data sources are from various stages.

It is a lot of complex in nature, in this way difficult to deal with.

The data continue changing time to time.

It is free from the impact, direction, or control of anybody.

This huge stockpiling of data requires enormous zone for real usage

4.2 The Evaluation Method

Data mining apparatuses are an extraordinary sort of programming that targets encouraging data mining occupations (Mikut, 2011). The correlation between programming items is a run of the mill assessment practice that has a place with the field of trial software engineering (Denning, 1981). We receive the standards of Design of Experiment (DoE) to manage assessment execution for choosing appropriate data mining instruments. **Figure 2** traces the means of DoE in eight stages. The depiction of each progression is as per the following:

1) Necessity Recognition: Identification of assessment prerequisites is the above all else task in DoE assessment strategy. These assessment prerequisites are important to grasp framework related issues just as the assessment reason. The specialists with earlier data of related issues conceptualize to record a lot of assessment necessity. From the start, the specialists utilize characteristic language to portray assessment necessities. They are then examined completely and mapped to necessities questions. This procedure decides the target of the appraisal procedure that is to address these necessity questions.
2) Highlight Identification: It institutionalizes the terms, ideas and their connections inside a framework space to decide a lot of highlights. The highlights incorporate both utilitarian and non-practical highlights. This progression uses the necessity inquiries to recognize important highlights for assessment.
3) Measurements and Benchmark Listing and Selection: Firstly, specialists research the pertinent measurements and benchmark and set up a rundown establishing them. At that point, the most proper ones are chosen dependent on the accessible assets close by, assessing the overhead of potential analyses, and passing judgment on the evaluator's capacities of working various benchmarks. The determination procedure is a urgent undertaking and assumes a fundamental job in assessment execution. When the measurement and benchmarks are picked, the determination of test components starts.
4) Exploratory Factors Listing and Selection: The trial components are the parameters or factors that influence the presentation highlights chose for assessment. Like the past advance (for example Measurements and benchmark posting and determination), this progression records a lot of potential up-and-comer test factors. Afterward, the choice procedure considers the competitor factors with most astounding effect and significance.
5) Trial Design: According to DoE, the following stage after cautious determination of the measurements, benchmarks, and test elements is to configuration tests for assessment. Before all else, basic tests are structured dependent on the pilot preliminaries. Later these analyses are adjusted for progressively complex plans utilizing DOE procedures. The progression yields exploratory guidelines, trial plans, and driving benchmarks. They are then used to actualize these tests.
6) Examination Implementation: This progression is to complete arrangement of analyses. For example, watching the conduct of the framework by continuously expanding the estimation of a determination factor and taking different perusing for each worth. The outcomes got from the execution step are pushed ahead for investigation.
7) Examination Analysis: In this progression, evaluators understand the outcomes and think about various frameworks on both useful and

non-practical grounds.

5. Experimental Implementation

The assessment investigations were actualized following the full factorial structure. The exploratory condition is steady for running both anticipating and grouping employments on RapidMiner and KNIME. On account of the determining employment kept running by RapidMiner, the work process incorporates three unique stages: 1) The primary stage peruses data from the CSV records, and afterward passes the data to the following stage. 2) The preprocessing stage readies the data and concentrates the preparation and testing sets from the whole data collection. 3) The last stage is made out of three noteworthy procedures, in particular data handling, approval, and plotting.

The readied preparing data collection is encouraged to the anticipating models, and the testing dataset is utilized to approve the estimating results dependent on the approval measurements, for example, root-mean-square error (RMSE) and mean outright error (MAE) (Chai, 2014). This stage additionally plots the outcomes with the end goal of perception. Alongside the evolving outstanding burdens, RapidMiner's execution time of running the anticipating occupation shifts going from around 29 seconds to 107 seconds, overall, for the data sizes from 1-month to 2-year.

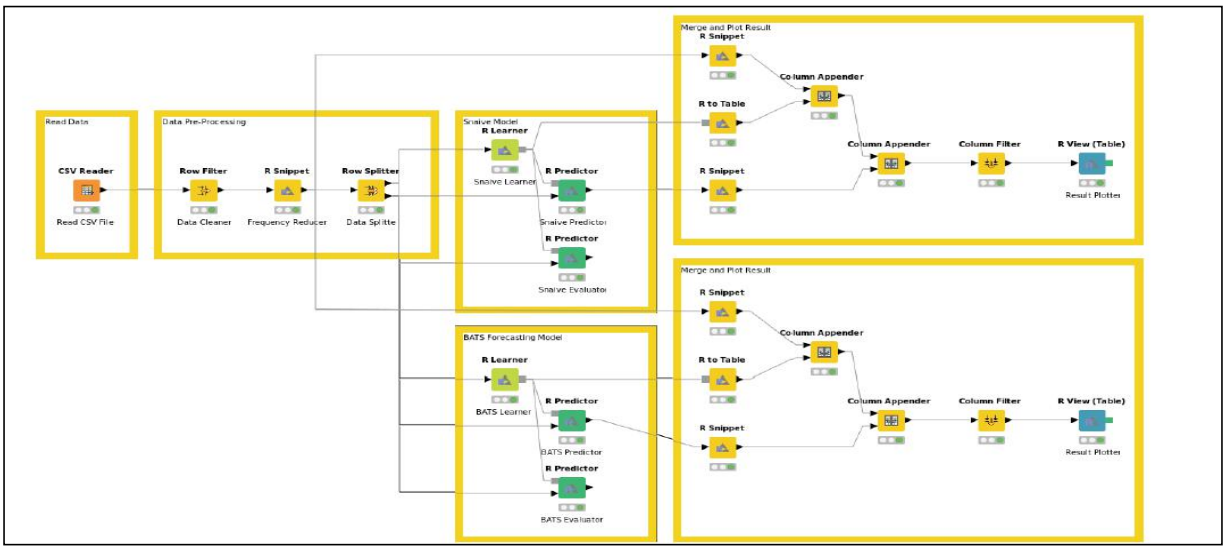

Figure 1. Forecasting job workflow implementation in KNIME.

6. Forecast to the Future

There are many future important challenges in Big Data management and analytics, that arise from the nature of data: large, diverse, and evolving. These are some of the challenges that researchers and practitioners will have to deal during the next years: 1) Analytics Architecture. It is not clear yet how an optimal architecture of an analytics systems should be to deal with historic data and with real-time data at the same time. An interesting proposal is the Lambda architecture of Nathan Marz. The Lambda Architecture solves the problem of computing arbitrary functions on arbitrary data in realtime by decomposing the problem into three layers: the batch layer, the serving layer, and the speed layer. It combines in the same system Hadoop for the batch layer, and Storm for the speed layer. The properties of the system are: robust and fault tolerant, scalable, general, extensible, allows ad hoc queries, minimal maintenance, and debuggable. 2) Statistical significance. It is important to achieve significant statistical results, and not be fooled by randomness. As Efron explains in his book about Large Scale Inference, it is easy to go wrong with huge data sets and thousands of questions to answer at once. 3) Distributed mining. Many data mining techniques are not trivial to paralyze. To have distributed versions of some methods, a lot of research is needed with practical and theoretical analysis to provide new methods. 4) Time evolving data. Data may be evolving over time, so it is important that the Big Data mining techniques should be able to adapt and in some cases to detect change first. For example, the data

stream mining field has very powerful techniques for this task. 5) Compression: Dealing with Big Data, the quantity of space needed to store it is very relevant. There are two main approaches: compression where we don't loose anything, or sampling where we choose what is the data that is more representative. Using compression, we may take more time and less space, so we can consider it as a transformation from time to space. Using sampling, we are loosing information, but the gains in space may be in orders of magnitude. For example Feldman *et al*. use coresets to reduce the complexity of Big Data problems. Coresets are small sets that provably approximate the original data for a given problem. Using merge-reduce the small sets can then be used for solving hard machine learning problems in parallel. 6) Visualization. A main task of Big Data analysis is how to visualize the results. As the data is so big, it is very difficult to find user-friendly visualizations. New techniques, and frameworks to tell and show stories will be needed, as for example the photographs, infographics and essays in the beautiful book "The Human Face of Big Data". 7) Hidden Big Data. Large quantities of useful data are getting lost since new data is largely untagged filebased and unstructured data. The 2012 IDC study on Big Data explains that in 2012, 23% (643 exabytes) of the digital universe would be useful for Big Data if tagged and analyzed. However, currently only 3% of the potentially useful data is tagged, and even less is analyzed.

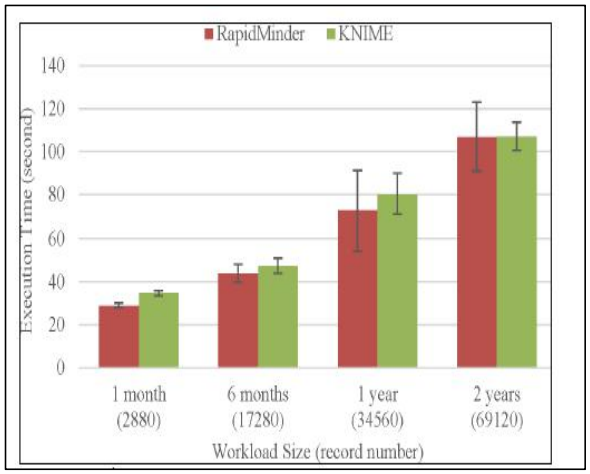

Figure 2. Average execution time of the forecasting job.

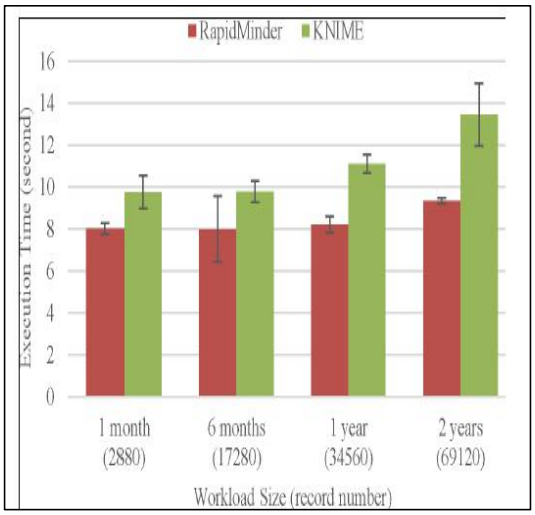

Figure 3. Average execution time of the clustering job.

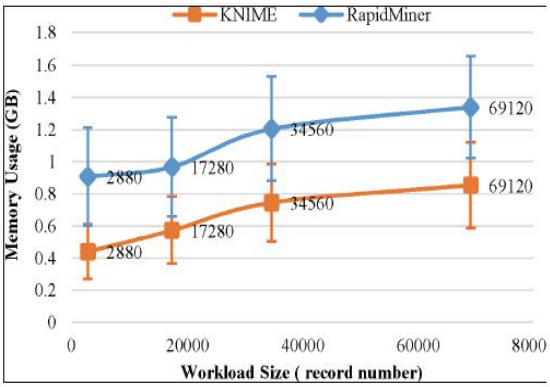

Figure 4. Average memory usage of the forecasting job against different sizes of workloads.

7. Experimental Analysis

7.1 Analysis of Quantitative Features

In term of Efficiency, we plot the ROC curve for the outlier detection techniques as shown in **Figure 5**. The ROC curve of the technique Local Outlier Factor (LOF) shows the best performance among the evaluated detection techniques. It shows the maximum TPR of 0.7 and minimum FPR of 0.15 approximately.

In contrast, OPTICS and DBSCAN have an almost same area under the curve and shows similar behavior according to the analysis. In terms of the Execution Time and the Memory Usage, we employ Pareto Chart to visualize the effects of the different experimental factors and their combinations. Two factors are plotted in the charts. Factor A represents Technique; Factor B represents Data Size, and Factor AB represents the

combinations of the two factors. We compare LOF with DBSCAN and OPTICS. From **Figure 6** and **Figure 7**, the analysis.

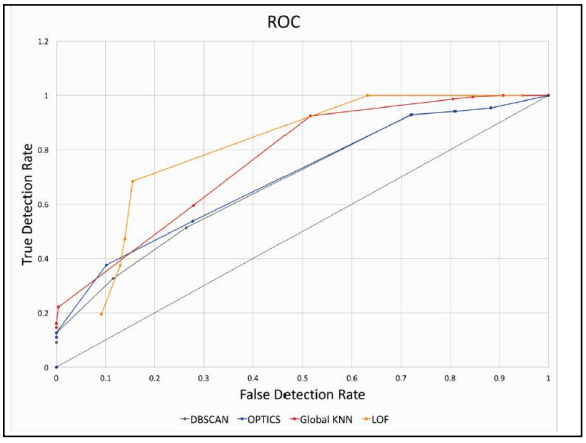

Figure 5. Efficiency analysis result.

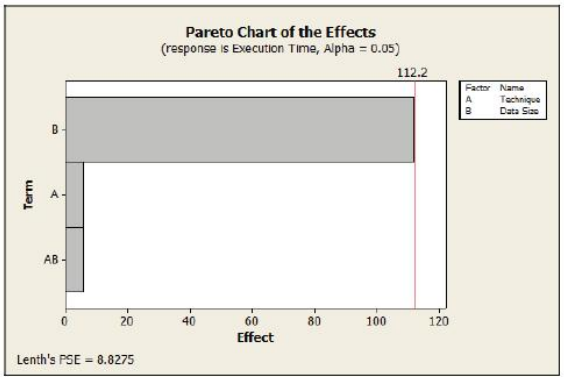

Figure 6. LOF vs DBSCAN (execution time).

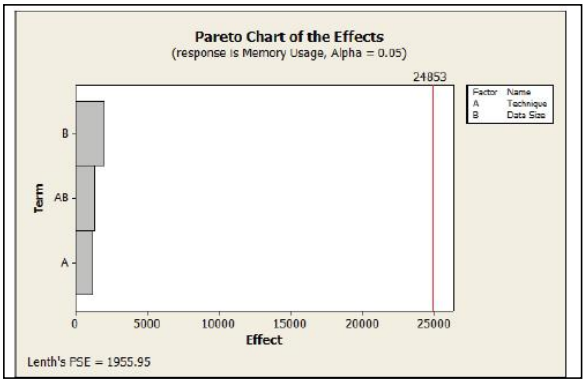

Figure 7. LOF vs DBSCAN (memory usage).

7.2 Analysis of Qualitative Features

In the case of outlier detection techniques, the quality attributes are used as selection criteria. The selection process is a subtask and forms a basis of quantitative analysis. The outlier detection technique that complies all these qualitative criteria will be assessed. Each qualitative criterion is analyzed independently (i.e. without considering its dependency on other features) for each outlier detection technique. The study results are illustrated in.

Distribution Independent: Most of the statistical outliers' detection techniques, such as Grubb's test and Chi-Square test, are distribution dependent. These statistical approaches rely on the underlying data distribution, and most of them are only applicable on a normally distributed dataset. In our case, these techniques are inapplicable as the base stations or eNodeBs generates data with distinct data distributions.

After scrutinizing the outlier detection methods in consideration, we found that Grubb's Test, Chi Square Test, Kernel Density Estimation (KDE), General Extreme Student Estimate are distribution dependent whereas the remaining techniques are independent of underlying data distribution.

Non-Parametric: Similarly, distribution independent methods are also non-parametric. Therefore, based on our analysis Grubb's Test, Chi Square Test, Kernel Density Estimation (KDE), General Extreme Student Estimate techniques are parametric, and remaining detection techniques show non-parametric behavior.

No Explicit Training: Telecommunication companies generate a large volume of data. The large volume makes training data preparation a tedious task. Therefore, outlier detection technique should not need any prior data preparation. Clustering and classification detection techniques such as One-class support vector machine (SVM), Bayesian naive, Random forest, K-Means, Self-Organized Mapping (SOM) are effective methodologies for outlier detection but need explicit training. Training these techniques require preparation of training data not feasible in our case. However, outlier detection techniques such as Grubb's Test, Kernel Density Estimation (KDE), DBSCAN, OPTICS, Global KNN, Local Outlier Factor (LOF) do not require any training data, thereby, they are more suitable for analysis.

Multivariate Data: Statistical outlier detection techniques i.e. (Grubb's Test, Chi Square Test, General Extreme Student Estimate techniques) are only applicable on a univariate dataset. Whereas, Kernel density estimation (KDE) has an implementation for both univariate and multivariate datasets. To the best of our knowledge, most of the machine learning algorithms

apply to both univariate and multivariate data. However, they tend to perform better on multivariate data than univariate data.

8. Conclusion

Huge Data is worried about the immense measure of data that are constantly developing, other than their exceptional speed that should be managed in an opportune way. Data mining can be utilized to find covered up, obscure, however valuable learning from huge, fuzzy, uproarious, fragmented, and irregular data. In this paper, we present our structure of test to assess data mining instruments and exception identification approaches into two diverse contextual analyses. The point of this assessment is to watch them both quantitatively and subjectively when running data logical occupations. Enormous Data investigation necessitates that circulated mining of data streams ought to be performed continuously. Much research is required in down to earth and hypothetical investigation to give new techniques to appropriated data mining with huge data streams. Since the data are constantly generated and accumulated, the workload size becomes one experimental factor in both the cases during the period of time when datasets are collected. The evaluation could become complicated because of the combination of factors. Therefore, applying DOE principles to our evaluation study clearly make this practice in a structured eight-step procedure.

References

1. B. Thakur, M. Mann. Data mining for big data: A review. International Journal of Advanced Research in Computer Science and Software Engineering 2014; 4(5): 469-473.
2. R. Vrbić. Data mining and cloud computing. Journal of Data Technology & Applications 2012; 2(2): 75-87.
3. V. Nekvapil. Cloud computing in data mining – a survey. Journal of Systems Integration 2015; (1): 12-23.
4. A. Bifet. Mining Big Data in Real Time. Informatica 2013; 37: 15 – 20.
5. G. Krempl, I. Zliobaite, D. B. Nski, *et al*. Open challenges for data stream mining research. ACM SIGKDD Explorations 2013; 16(1): 1-10.
6. D.-H. Tran, M. M. Gaber, K.-U. Sattler. Change detection in streaming data in the era of big data: models and issues. ACM SIGKDD Explorations 2014; 16(1): 30-38.
7. W. Fan, A. Bifet. Mining big data: Current status, and forecast to the future. ACM SIGKDD Explorations 2012; 14(2): 1-5.
8. Y. Demchenko, P. Grosso, C. D. Laat, *et al*. Addressing big data issues in scientific data infrastructure. 2013 International Conference on Collaboration Technologies and Systems (CTS), 20-24 May 2013, San Diego, CA, USA, pp. 48-55, 2013.