International Journal of
INTELLIGENT SYSTEMS AND APPLICATIONS IN ENGINEERING
ISSN:2147-6799 www.ijisae.org Original Research Paper

Reinforcement mSVM: An Efficient Clustering and Classification Approach using reinforcement and supervised Techniques

Dr. Satish S. Banait, Dr. S. S. Sane, [2] Dr. Dipak Bage , [3] Prof.A.R.Ugale [4]

Submitted: 06/06/2022 **Accepted:** 10/09/2022

Abstract: Data mining as well as big data analytics represent approaches for analysing and extracting useful secret data. Although big data is complicated and large in volume, conventional methods to interpretation and retrieval do not function well. Data clustering is a common data mining approach that divides nodes into categories and makes it possible to retrieve features out of these groups. Conventional clustering techniques, including such k-means clustering as well as hierarchical clustering, are inefficient because the reliability of the groups they generate is harmed. As a result, an efficient and relatively extensible clustering technique is required. In this paper we proposed novel similarity-based clustering techniques on large unstructured transaction dataset. The HDFS file system log data has collected from real time Virtual Machine's (VM's) and generates the clusters, using reinforcement learning technique. Initially data has collected from various VM's and proposed dimensionality reduction technique has used for data reduction. The Q-learning based reinforcement learning algorithm has applied on generated event. The activation function calculates the current weight for each transaction according to reward and penalty. Finally, the threshold-based entropy function generates a final cluster. After the clustering process modified Support vector Machine (mSVM) as supervised classifier has applied on entire label data. In the extensive experimental analysis, we evaluate proposed model performance with existing techniques. The proposed clustering and classification method beats the comparable models in terms of average operating time and average clustered error, according to tests conducted on actual, synthetic, and automatically created datasets.

Keywords: Big data mining, efficient clustering methodology, Unsupervised Learning Technique, Data mining, k-means

1

Introduction

Big data is producing a lot of attention in the marketplace today and data is fast expanding from gigs to battalions, exabyte, and zettabytes [1]. Big data has such huge data demands that earlier data storage and processing technologies, such as Database Systems (DBS), Relational Database Management Systems (RDBMS), and others, are currently struggling to meet the needs [2]. Big data entails extraordinarily huge datasets, making it impossible for frequently used programs to handle and analyse it in the timeframe which is necessary. To accommodate this burden, parallelized software that runs across multiple machines is now required [4].

Data gathering and preservation are now considerably easier and less costly than they have ever been thanks to continuous scientific and technological advances. This resulted in the creation of big data in research, administration, and business, which needed to be analyzed or categorized in order to extract meaningful data. If we examine the outcomes of a search engine for a certain request, for instance, the user must filter through vast lists to get the optimal output. However, whether there is millions of internet pages offered as methods for solving query, this work might be quite tough for the customer. Clustering is described as the unsupervised classification process of data objects or views, i.e. the information have still not been divided into any groups and thus have no class labels. Among the most crucial processes in information extraction is clustering [12]. To locate the relevant and undiscovered groups of similarities, clustering methods are applied. Clustering is a technique for dividing items into clusters of things that are similar. Disparate objects are grouped together in different clusters. A data item may apply to a particular cluster or perhaps more than a cluster, based on the criterion selected. Assume a store database that contained data on the things that customers have bought. Clustering divides the customers into groups based on their purchasing habits. When we organize things into

ssbanait@kkwagh.edu.in1 , ssane@kkwagh.edu.in2, ddbage@kkwagh.edu.in3, ar.ugale@gmail.com
Department of Computer Engineering, KKWIEER, Nashik, SPPU Pune, India1, 2, 3,
Department of Computer Engineering,MET's Institute of Engineering,Nashik,SPPU,India4

clusters, we get clarity at the expense of certain data. The decision on which clustering algorithm to use is crucial.

As a consequence, clustering algorithms can be particularly beneficial in combining the strongly linked methods to solve query and presenting the outcomes in terms of groups, allowing irrelevant articles to be ignored even without looking at them. The primary principle underlying clustering any set of data is to detect underlying statistical regularities and understand it as a groups or clusters, in which the data items within every group should have a high level of intra-cluster resemblance, but the resemblance across groups should indeed be minimized. Clustering is used in a variety of contexts, including • News articles: Sorting everyday media stories into categories such as athletics, features, economy, and fitness, among others. • Classification of digital materials (WWW): Web search outcomes can be grouped as per the degree of correlation for the specified query. • Market exploration: Given a massive dataset including each single customer's historical purchase information, identifying groupings of customers that behave similarly. • Research work: Gathering big quantities of information from devices on a daily basis is pointless unless certain conclusions are reached. Detecting and categorizing essential relationships in gathered information could lead to results of the study suggest. • Disaster studies: grouping observed seismic meccas to determine risky zones. The primary issues with classic clustering methods are dealing with broadness and sustainability as data sizes expand rapidly. The complexity of data grows as it grows in size, wreaking havoc on the throughput and storage needs of huge applications.

Clustering is the partitioning of information into groupings of products in relation. Every group (= cluster) is made up of things that are comparable to one another but not to items from those other clusters. Clustering or grouping can be thought of as unsupervised learning of ideas from the standpoint of Machine Learning (ML) [5]. The following is a easy, technical, theoretical explanation of clustering, as given in [6]: Let X Rmn symbolize a collection of m locations xi in Rn as a group of data elements. The idea is to divide X across K groups Ck so that information from the very same unit is more "similar" than information from other categories. A cluster is a collection of K groupings. The procedure produces an injective map XC of data objects Xi to groups Ck as a consequence. In latest days, a dramatic shift in online usage and an enhancement in communication have led in the storage of huge amounts of data in databases. This need prompted many academics to consider methods for finding data and classifying it such that valuable data might be obtained. Clustering can be accomplished in a number of ways, with distinct sorts of clusters produced by every approach. Some need human input, such as the number of nodes to be produced, while others depend on the kind and volume of information provided. The invention of density-based as well as grid-based clustering algorithms was the most significant advancements.

Clustering is a machine-learning-based data mining method for dividing groupings of conceptual items into categories of related things. Clustering aids in the division of data into groups. These groups are made up of data items that have a high inter-similarity but a low intra-similarity. The following factors are used to classify clustering techniques [1]: • The kind of input • A condition for grouping items based on their resemblance • Ideas that underpin clustering methodological approaches Partitioning approaches, modular methods, model-based methodologies, density-based techniques, and grid-based strategies are the five different categories of clustering techniques.

Centroid-based clustering – Unlike the hierarchical clustering described below, the most extensively used centroid-based classification algorithm is k-means. The efficiency of centroid-based techniques is limited by their sensitivity to boundary conditions and anomalies. Since k-means is an effective, successful, and easy clustering method, it is the subject of this study.

Density based clustering – Density based clustering joins areas with high instance concentration into clusters. As long as concentrated regions can be linked, arbitrarily chosen dispersion is possible. These techniques struggle with information with a wide range of concentrations and sizes. Furthermore, these techniques are not designed to allocate outliers to groups.

Hierarchical Clustering Techniques — Hierarchical clustering divides a set of data items into layered clusters or a forest architecture. Although the outcomes of K-means grouping and hierarchical clustering algorithms may appear similar at times, their methods are different. As opposed to the K-Means technique, there is no need to accurately predict the number of nodes.

Grid based clustering – It is the method that analyzes an item space and quantizes everything into a limited cell count, establishing a regular grid. The function then does all of the procedures on the grid system. The key feature of this technique is that it has a quicker turnaround time that is different for different of items and only based on the quantity of cells within every dimension of the quantum space. Grid-based approaches divide the whole issue area into cells that used a single standard grid mesh, as well as the data items within each cell are recorded by the grid to use a range of numerical characteristics from the items. Rather than the data,

segmentation is done on the grid cells. The computation time can be greatly enhanced because the mesh density is much lower than the input of data items.

Clustering based on distributions- This clustering method implies that data is made up of statistics, such as Gaussians. The data is clustered into three Gaussians using a dispersion technique. The likelihood that a location corresponds to the distribution diminishes as the radius out from distribution's center increases. These lines depict the likelihood drop. One must use a different methodology if they don't know what sort of distribution of given information.

We arrive at a point where we could all draw a consensus regarding why traditional clustering techniques have difficulty with big databases after examining all of the main classifications of clustering techniques. We see that if we really want to lower the computing implementation time, we must abandon the use of distance space. Distance-space functions-based techniques appear to have more scaling issues compared to their vector-space counterparts. It takes O to calculate and store the connection among all possible combinations of n things (n2). When the readings are of a maximum variance, determining the difference between two items might be costly. There is no specified procedure for selecting a cluster's "central," and doing so ad hoc increases to the computation time. Vector-space approaches obviously have certain benefits above distance-based techniques. We could describe the vector-space with statistics derived from items in the group if we really want to enforce a statistical method on it. Vector-space frameworks' capacity to create "reliable" approximations of every group can be leveraged to reduce memory and computation expenses. Clustering techniques' scalability, the efficacy of strategies for grouping intricate forms and kinds of analysis, elevated clustering approaches, and techniques for grouping combined numerical and categorical in huge databases have all been studied.

The operational difficulties of every clustering approach for managing big data are one of their concerns. Decentralized network ideas could potentially enhance the accuracy of existing approaches. More research towards the use of groups is needed to confirm the influence of multi-view clustering methods on Big Data. Numerous studies should address some of the problems, such as the shortcomings of current clustering techniques in handling randomly established knowledge distribution of sets of data, quantitative evaluation of the benchmark of clump outcomes, and preventing the need to clarify input variables for a clustering method, among others.

Big data necessitates approaches for extracting insights from large, diverse, and complicated information. Limitations in data gathering, fulfilling necessity speed, resolving information quality, coping with abnormalities, spreading big data, as well as big data analysis are among the issues of big data processing. ML, connection approaches, support vector machines, as well as grouping are just a few of the strategies that have been developed to deal with massive data collections. Clustering aids in the identification of clients who share a common customer history and behavior. Clustering analysis is widely employed in a variety of fields (e.g., market testing, pattern identification, information and image analysis). Clustering can also assist marketers in recognizing various groupings of customers within their database. Customer groups, for instance, can be identified by purchasing behaviors. The suggested effort seeks to collect and evaluate various big data sets, find possible current clustering techniques, investigate, execute, and assess this clustering technique, and create detailed an efficient Big Data clustering technique. Conclude that the data by testing the developed method on available information sets and comparing and analyzing the effectiveness of the proposed scheme with that of current programs.

In this study, we present a novel and efficient clustering method for dealing with large amounts of data. The remainder of this work is arranged in the following manner. In Section 2, current clustering is detailed, as well as its benefits. Section 3 describes the proposed clustering method. The experimental procedure is described in Section 4, and the outcomes of trials utilizing the suggested technique are covered in Section 5. Lastly, in Section 6, we review our findings and offer recommendations for further research.

Literature Survey

Mohamed Aymen Ben HajKacem et al. [1] created a unique kprototypes clustering algorithm based on Spark in 2017. While coping with huge amounts of different datasets, the k-prototype technique has two major drawbacks: runtime as well as storage usage. The usage of Apache Spark to parallelize the k-prototypes approach created a simple, clear, and sustainable environment. The test findings suggest that new approach is extensible and can outperform current k-prototypes techniques. Employing dimensionality reduction methods, the system can manage vast amounts of various datasets with a huge number of parameters. A further area of scientific investigations is to use the given KP-S approach to fraudulent activities in the telecommunications field. Ms. Tejaswini U. Mane et al. [2] described big data and its qualities, as well as its challenges and opportunities, in 2017. Today, the word "Big Data" is more and more widely used. Big data refers to a large volume of diverse data that is growing at a rapid rate with the passage of time. So, rather than simply data storage, it is essential to

validate it and retrieve some useful info out it using data mining algorithms such as clustering and classification. Because Big Data covers a wide range of eras, it was chose to start with the health industry. There are a variety of ailments on which to research, learn things, or anticipate, but cardiovascular disease was the only one examined. Heart disease is one of the most dangerous diseases, with a higher death rate than other illnesses, as per the World Health Organization. As a result, in the study work, it was chosen to employ cardiovascular disease as a big data strategy by using Hadoop MapReduce framework. In the hybrid technique, Enhanced K-Means were utilized for clustering and indeed the decision tree algorithm ID3 was employed for categorization. Obtaining an independent advice is becoming increasingly common; the platform is extremely useful for assisting in prediction based on certain characteristics such as chest discomfort, triglycerides, age, resting Bp, Thalac, and others. Medical decision making will be enhanced of this technology. It will also have an effect on the therapy process's development. As a result, it will be extremely valuable in predicting cardiovascular problems.

Big Data is a term used to describe a large volume of data that is being investigated via IOT (Internet of Things) of many resources like sensing devices, Facebook feeds, and web technologies. Traditional tools and techniques are incapable of handling Big Data. Online communications are increasingly dominated by social networking sites. Big Data mining is done in order to generate profit from large amounts of data in addition to provide superior insights through the use of suitable measures. Association Large datasets do not fit in main memory, therefore rule mining and frequent itemset mine are prominent data mining approaches that require the complete dataset to be loaded into primary memory for computation. To get through this issue, MapReduce is used for concurrent Big Data processing. It has properties like great scaling and resilience, which makes it easier to deal with enormous datasets. Sheela Gole et al. [3] present a new techno, ClustBigFIM, that performs on the Mapreduce model for mining large datasets. ClustBigFIM is an altered BigFIM methodology that provides adaptability and pace in extracting purposeful data from huge datasets in the manner of correlations, emerging and sequential patterns, correlations as well as other important data mining activities. JIAN YIN et al. [4] propose a different clustering approach for huge datasets with mixed features in 2005. It solves the problem of standard methods being unable to handle mixed characteristics in huge datasets. The technique pre-clusters datasets using CF*-tree and stores dense areas in leaf nodes, which enhances the system's runtime performance. The approach has a good throughput and therefore can fast and effectively cluster massive data with diverse properties.

An effective approach for clustering remote databases was developed by Ahmed Elgohary et al. [5] in 2011. Iterative optimization is used to develop an effective approach for clustering distributed databases in the manner of a Peer-to-Peer connection. The developed approach outperforms a newly enacted method based on a distributed variant of the well-known K-Means technique, according to the findings presented in this study. The novel method is expected to observe widespread use in remote clustering scenarios wherein effective responses are needed. Yu-Fang Zhang et al. [6] introduced the enhanced K-means method as a remedy to manage large scale data in 2003, claiming that it can intentionally pick initial clustering centers, lower sensitivity to isolated points, prevent dissevering big clusters, and resolve data defluxion to some extent induced by disparity in data fragmentation due to multi-sampling adaptation. Sheng-Yi-Jiang et al. [7] investigate the topic of grouping mixed extracted features in huge databases in 2004. To begin, a distance specification for any sort of information to computer distance of two things, length among an object and a cluster was supplied. The limit of variable was then determined using a simple method. An efficient clustering technique was presented depending on the novel proximity, which can deliver great clustering outcomes and has strong extensibility. The technique can be used to group data streams. The outcome indicates that the length described in the research is capable of accurately measuring the object disparity, and the approach for determining the threshold is straightforward and encouraging.

Sheng-Yi Jiang et al. [8] presented a 2-stage hybrid clustering technique in 2009, combining the one-pass clustering technique and the DBSCAN clustering method. Not only does the novel clustering technique process information with categorical variables, but it can also find groups of any form. It has a temporal difficulty that is almost linear with corpora size and may be applied to huge datasets. The test data on different datasets reveal that the provided technique is more efficient than some other well-known algorithms in terms of information type, adaptability, performance, and clustering accuracy. Hui Zhang et al. [9] established a novel Key-Feature Clustering (KFC) technique in 2006 that aggregates search terms from the outcomes as key characteristics before clustering the articles based on some of these clustered major features. Finally, the paper summarizes and evaluates the findings of tests performed to evaluate and confirm the method. For the needed computing time as well as objective function, Rasim Alguliyev et al. [10] contrast the standard k-means

clustering technique to the suggested Batch Clustering (BC) approach in 2020. The BC technique is intended for batch clustering of huge data sets while maintaining performance and accuracy. Numerous investigations show that when matched to the k-means technique, the Batch Clustering approach for large databases is more effective in terms of computing capabilities, information storage, and clustering. Investigations were performed out utilizing a data collection of 2 million 2 D data points.

Carlos Ordonez et al. [11] presented an ER-Flow combination diagram based on contemporary UML syntax in 2020 to aid investigators in information pre-processing and discovery. In order to make the ER model as simple as possible, interconnections were expanded with an axis to indicate processing activity, and structures resulting from pre-processing were tagged with scores and alteration tags. It demonstrates how a revolutionary ER-Flow diagram can assist users in navigating huge amounts of data just at metadata stage, offering a combined view of information and source code, as well as other useful features. JIAN YIN et al. [12] tackle the issue of privacy preservation in big data analytics in 2015. Many interesting rules are generated from huge data using machine learning approach in a variety of industries. However, data privacy must be safeguarded. In many cases, data secrecy may be required before information can be obtained. A technique for proving safe data secrecy was described, as well as several methods to enhance the system's efficiency. Luo Xiaofeng et al. [13] presented an objective (big data info service quality), 3 chains (data value chain, IT value chain, and data assurance value chain), and 5 responsibilities for the big data reference system framework in 2020. (Supplier of big data applications, big data frameworks, and big data information security) Many of the system's properties were represented in a 1-3-5 framework of information supplier and data consumer. The flaws in ISO / IEC TS 25011:2017 in the area of big data information quality of service have been resolved. The primary big data framework was then suggested. When comparing the big data reference architectural system suggested in the research to several other popular method, the conclusion is made that some other common models might be effectively overtaken.

Ikbal Taleb et al. [14] addressed the establishment of Data Quality Rules (DQR) post performance analysis and before Big Data pre-processing in 2017. The DQR discovery paradigm was presented as a way to improve and precisely concentrate pre-processing tasks relying on quality criteria. To automate the DQR creation, a collection of pre-processing processes linked therewith Data Quality Dimensions (DQDs) were created. To reduce multiple pass pre-processing operations and remove redundant standards, rules optimization was performed to verified rules. Tests demonstrated that using the identified and improved DQRs on data resulted in higher quality rates. David Becker et al. [15] published a paper in 2015 that looked at numerous aspects that affect Big Data Effectiveness on numerous scales, encompassing gathering, computation, and preservation. Although perhaps not surprising, the article's core results reveal that the constraints and challenges of processing Big Data while keeping its validity are the fundamental variables that influence it. These issues outweigh those about the information's origin, storage, and methods for preparing, manipulating, and storing the information. Including all data analytics concerns, information quality is fundamental. The "reality about Big Data," according to the study's conclusions, is there are so many completely novel DQ challenges in Big Data analytics programs. However, several DQ concerns show return-to-scale effects, which will become somewhat prominent in Big Data analytics. The reliability of Big Data varies from a sort of Big Data to the next, including from one Big Data technology to the next. Ernesto Damiani et al. [16] offered a characterization of the Big Data Leak threat (as opposed to a data breach) in 2015, as well as its position as an element of exposure vulnerability. It then went over how a Knowing, Recognize, Contain, as well as Recovering methodology may be utilized to build Big Data security standards for limiting exposure hazards associated with Big Data analyses.

In 2020, Bhagyashri S. Gandhi et al. [17] published a fresh review on the concept of big data that consists of massive traditional heterogeneous dataset that is constantly evolving and complex in character and obtained from multiple sources. Conventional database methodologies are incapable of recognizing, evaluating, and handling these massive data warehouses. Because traditional data processing technologies are unable to provide adequate assistance, the accessibility and gathering of information are the 2 key challenges that must be addressed. To combat these characteristics of big data, a mixed supervised as well as unsupervised learning strategy with an ensemble method in a ubiquitous computing environment is suggested for lowering high - dimension data to enhance overall effectiveness. This document describes the various ways of dealing with massive data that have been presented. Disha D N et al. [18] published a study in 2016. The Map Reduce Architecture is used to perform data extraction and analytics utilizing Twitter as a large data repository. Threads of tweets are whittled down to a tolerable size. As a result, data categorization, predictive evaluation, and data clustering can be accomplished quickly. Few basic evaluation is done out by using Naive Bayes (NB) technique for sentiment analysis of user ratings in order to estimate the performance of the

classifier using the binary check. The study yielded a decent outcome and a confusion matrix for the user ratings, which were divided into favourable and unfavourable categories. The execution leads to the Bernoulli hypothesis, and it may run a binomial test to calculate a confidence interval for the outcomes. The population percentage mean would be successful is between 61.28 percentage and 63.13 percentage, according to the 95 percentage confidence range.

Zakaria Gheid et al. [19] presented sk-means, a highly innovative and confidentiality k-means clustering technique based on the safe and simple multiparty additive protocol Π-sum. It demonstrated the safety of sk-means and Π-sum, along with their elegance and effectiveness when contrasted to certain other assertions, through various tests. The preliminary findings of the test indicate that the method is more suitable for big data properties and extends to large databases. Maryam Abdullah et al. [20] used fuzzy clustering methods in a broad spectrum of clinical applications earlier. The FCM method was primarily used for picture segmentation. Furthermore, the effectiveness of the Fuzzy C-mean and Gustafson Kessel Clustering techniques on health data supplied as from UCI-ML Repository was explored in this research; specifically, the WBC database was used in the proposed approach. Numerous studies were carried out to evaluate the techniques' effectiveness and to determine the appropriate specifications for FCM and GK. When m = 2, the findings show that FCM and GK performed the best. The findings indicate that FCM performed better, with the optimal outcomes of around 95.25 percent accuracy rate, compared to 91.7 percent classification results for GK. Whenever m was set to 2, FCM obtained decreased categorization error, as evidenced by comparisons to previous research. Additionally, the data suggest that FCM takes somewhat less time to execute over GK.

In 2020, Doaa.Sayed et al. [21] proposed CluStream, an efficient information streams technique for clustering big information streams. The presented method improves on the classic CluStream by using a highly effective clustering algorithm and creating groups more than a sliding window. CluStream is not suitable for usage with huge data as all information is held and analyzed, and therefore no past data is deleted. SCluStream contains an expiry stage that removes expired pictures from the window size, preserving storage and allowing the ultimate cluster to be built from even the most latest information.

Louis Y. Y Lu et al. [22] proposed a novel method for detecting major research ideas in the big data domain in 2016. The reference chain among 3898 publications was established using Scopus as well as Web of Science, and edge-betweenness segmentation was used to determine the primary research topics in the big data domain. The primary research topics disclose the current research issues, while the development curves suggest future directions. The findings are a fantastic resource for people hoping to learn more about the evolution of big data and are extremely useful in filling in the gaps in the existing literature. Data mining techniques based on large data occur in the execution effectiveness, algorithm parallel processing, and framework ease-of-use, according to Xinxin Huang et al. [23]. To resolve these concerns, the study investigates the data mining algorithm in the context of big data; it uses Spark as the primary engine and uses Spark's memory computation function to perform parallel computing of numerous classic data mining techniques. For a smooth implementation, the data mining technique can execute in simultaneously in a cluster setting using big data. The data mining method then employs a layered design and creates a comprehensive big data mining framework using the platform layering technique. The concurrent machine learning technique has been enhanced as a result of this article's work; it allows people to have more options in the process of data exploration; and it reduces the platform's operation sophistication, allowing employees to concentrate on methodologies and industry rather than the fundamental project implementation.

Galina Chernyshova et al. [24] looked into the advantages of using the k-means technique in text categorization. M. Omair Shafiq et al. [25] developed a scheme that incorporates a revised framework for semantic logs and an event categorization method that expands k-means clustering over MapReduce and allows it to group and analyze logs modeled utilizing suggested improved semantic logging paradigm at a massive scale. The suggested approach was created in such a way that it can run in simultaneously on every node density and is unaffected by the amount of log event clusters to be detected. The recommended optimizer difficulty has also been examined, and has been determined that it is capable of handling logs on a big scale. According to Dr. Anu Saini et al. [26], the rapid increase in the rate of data collected from different origins has necessitated the efficiency of the existing algorithms for analyzing massive datasets swiftly. K-Means and other clustering methods are important for analyzing datasets. To obtain best outcomes, K-Means relies heavily on good initiation. The K-means++ activation approach provides an answer by giving the K-Means technique an initial set of centers. Nevertheless, because of its underlying sequential structure, it has a number of drawbacks when used to huge corpora. It takes k cycles to discover k centers, for example. The goal of this study is to describe a method that aims to solve the flaws of prior methods. The research created a strategy for identifying a suitable

initial seeding in very little time, allowing for rapid and precise cluster analysis across big datasets. The different advantages as well as constraints of MapReduce have been brought out too and identified by Seema Maitrey et al. [27]. MapReduce is a basic data analysis framework that offers strong throughput and low latency.Even for data warehousing, nevertheless, MapReduce is difficult to accept Database Management System (DBMS). Rather, MapReduce is planned to supplement DBMS by providing scalable and adaptable parallel processing for a variety of data analysis, including scientific information processing. However, for effective ramifications, performance, particularly I/O expenses of MapReduce, must be handled.

Dajung Lee et al. [28] suggested a reliable multilevel, real time data clustering approach with flexible hardware platforms that can be implemented in a heterogeneous environment. The methodology first uses streams of data to generate numerous sub clusters before applying an issue specific clustering technique to these sub clusters. Every sub cluster is made up of a series of centroids, each of which can be evaluated using just a distinct set of requirements. Every sub cluster unit receives continuous raw data and updates the centroids set as additional information is received. The very next stage is to group these estimated points, which is specified by the dataset attributes, by mapping centroids to groups. However unlike k-means technique, which has only one represented point for every cluster, the new technique allows each cluster to have several centroids. Suyash Mishra et al. [29] investigated ways for evaluating complex data in order to retrieve information from it. Furthermore, numerous methodologies and tools that were required to manage, analyze unorganized huge data efficiently, and improve the efficiency of complexity analysis were detailed. Ni Bin et al. [30] did extensive study on Internet of Things (IOT) big data clustering analysis techniques and practices in 2018, and created and constructed an information gathering processing and prototyping system that relies on the clustering prediction model.

In 2019, S. Dhanasekaran et al. [31] used improved map reduce approaches depending on k-means clustering for huge data analytics to improve map reduce and maximize memory on cloud corpora. Ankita Saldhi et al. [32] proposed in 2014 that a single DataNode can perform numerous parallel patterns of communication such as collect, filter, scan, and minimize. It also intends to achieve 2 parallel scan methodologies: (1) Hillis and Steele scan and (2) Blelloch scan. Saurabh Arora et al. [33] conducted a comprehensive review of modern clustering techniques for data mining on a wide scale. It has been discovered that noval procedures are still necessary for studying large data, as conventional methods are ineffective for assessing actual and internet streaming information. G. Anuradha et al. [34] proposed clustering and mining approaches for data streams in 2014. Internet logs can be used to monitor and mine changing web user habits via data stream analysis. Ensemble systems can be utilized in uses like Internet traffic, spam filtering, and Intrusion Detection (ID) when data reaches at a fast rate. The use of Genetic Inspired ML in bioinformatics and biomedicine is the subject of a lot of investigation. Each method has its own set of benefits and drawbacks. Because the quantity of big data is growing by the day, highly effective approaches for grouping and mining such massive data are urgently needed. Doaa. Sayed et al. [35] suggest CluStream, a CluStream upgrade, for monitoring clusters in 3 stages: active, expiring, and offline, in 2020. CluStream monitors groups over a sliding window in the initialization step to concentrate on the most current data to tackle the aforementioned data streams problems.

Charalampos Chelmis et al. [36] devised a new time series clustering model called on Hausdorff distance that effectively clusters structures using the distance measurement as well as information stashing method. The suggested scheme is scalable to huge data sets and is not limited to data on power usage. Ishwank Singh et al. [37] used the K-means clustering technique to groups of learners into distinct clusters in 2016. It'll also assist learners and educators in focusing on improvement initiatives by tracking student achievement. Spiliopoulos et al. [38] describe a MapReduce-based analytical solution for clustering large geographic data, such as that acquired from ships via the AIS, in 2017. Fadia Alaeddin et al. [39] provide state-of-the-art solutions based on adaptive approaches in 2020, which are designed to solve and solve challenges that arise while dealing with Big Data mining issues, particularly clustering, categorization, and feature extraction. Jungkyu Han et al. [40] suggested a fast k-means technique based on a statistical bootstrap technique in 2014. The developed methodology can achieve equivalent precision to the standard k-means technique when using the Lloyd approach on the original data, but with a substantially shorter runtime. Scalable Random Sampling using Iterative Optimization was suggested by Neha Bharill et al. [41] in 2016. SRSIO-FCM is a fuzzy c-Means. It is a robust version of RSIO-FCM that has been modified to address the issues of fuzzy clustering in Big Data.

Sunil Kumar et al. [42] used Hadoop MapReduce to validate the new hybrid clustering algorithm in 2019. Researchers evaluate clustering techniques utilizing NCDC weather information files when coping with huge data. They set out to discover the hottest day of a certain year by grouping data using several clustering techniques. Every method has disadvantages; for

example, K-means clustering method creates only just few groupings and necessitates pre-defining the number of nodes to be established because it is stable, so even though hierarchal clustering algorithm is constantly evolving and creates more groups than k-means, but somehow it necessitates numerous repetitions due to some kind of many intermix and divided choices. Due to these issues, the two algorithms are blended to get the benefits of both while ignoring the drawbacks. The highest number of clusters located from a file using the associated hybrid model, and the clusters created are of very high quality, leading in its most efficient findings. The suggested hybrid methodology results in a more efficient clustering method with improved accuracy, recall, as well as F-measure. Because the estimated maximum temperature level is the top limit temperature level, the outcome by the advanced hybrid clustering method is the most reliable. The hybrid algorithm generates the most groups and incorporates all data points in any of these groups. In the Mapreduce model, the Mapping output values are also optimized for this process. In many ways, the suggested method for effective big data grouping outperforms conventional clustering algorithms, although the hybrid clustering method has the drawback of requiring longer performing than either the k-means as well as hierarchical clustering techniques. Future studies are needed to see if alternate settings may be used to lower the processing duration.

Garima et al. [43] published their findings in 2019. Clustering is the method that divides a set of data into groups or categories called clusters so that data sets that are comparable are grouped together within one group. Because of the wide range of data types, clustering is very significant in data mining. This study examines the many classification methods required for data mining and compares clustering techniques such as DBSCAN, CLARA, K-Means clustering, CURE and others. As a result, this study covers the computational complexity of several clustering methods such as divisional clustering technique, hierarchical clustering method, density-based clustering strategy, as well as Grid based clustering algorithm. In essence, partitioned clustering depicts groups have used a template. Whenever the groups are of convex structure and same dimension and the number of nodes can be determined beforehand, partitioned clustering techniques are particularly beneficial. Hierarchical clustering methods are utilized due to the difficulty in forecasting the number of groups in before. It split the input into clusters, which are various stages of division. These methods are quite efficient in mining, but the idea of producing huge datasets is excessive. Because they can quickly recognize noise and cope with groups of any size, density-based clustering algorithms are highly effective in extracting huge datasets.

Proposed System Design

The below Figure 1 describes a dimensionality prediction based cluster generation and classification using supervised machine learning approach. In first we generate the clusters with dimensionality reduction using optimization techniques then classification has applied on entire data.

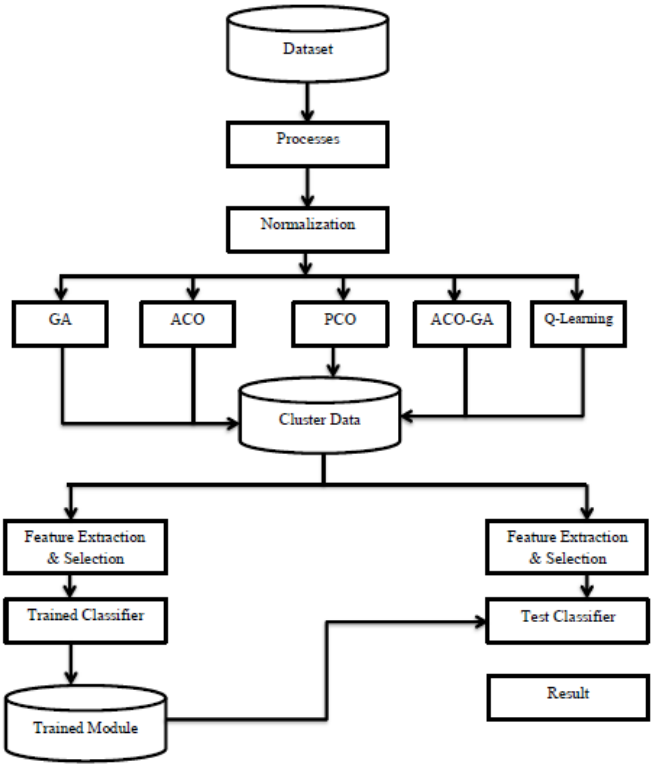

Figure 1: proposed system architecture for dimensionality reduction and classification using machine learning

Data Pre-processing and Normalization - Various data cleaning strategies are utilized in this step, including handling null values handling, redundancies, and unnecessary data. Data can be inaccurately attributed and noisy. Finding and modifying (or removing) lost erroneous, inaccurate, or meaningless information and substituting, updating, or removing confidential material is required to clean and fix false or inaccurate data from the records and datasets. We balanced the data using consistent sampling techniques, and we filtered the standard dataset to remove instances that were wrongly classified.

Feature Extraction- Feature extraction is a crucial process in a pattern identification system. The algorithm supports traditional feature extraction methods such as TF-IDF, Co-relational co-occurrence, N-Gram, Bi-Gram,

etc. We used those conventional methods as well as new dependency features from the entire input data.

Feature Selection – Sometimes the retrieved features have a low density, and other times they have redundant features. As a result of such a feature set module's potential to overfit or provide a high error rate, it is necessary to optimize the complete feature set. During the feature selection phase, we used various quality criteria to gather excellent features, which guarantees practical module training.

Classification: After cluster generation f the entire data, we apply a supervised machine learning classifier to validate the accuracy of the proposed model. The modified Support Vector Machine (mSVM) classification algorithm has been used for classification. The accuracy archived by mSVM with various optimization functions has demonstrated in the result section in detail.

Algorithm Design

The above algorithm technique presents a solution that gets around the problem of classification being predicted, and it does so use the hybrid supervised classification method. The total values for the retrieved parameters are shown by the proposed classification technique. These values are calculated using a two of machine learning algorithms. The following methodology was used in order to construct the machine model with the combination of reinforcement learning based prediction method.

Q-Learning Cluster Generation

Step 1: Generate each event reward and penalty for similar grouping

Step 2: calculate the score of each item using below function

$$fx = \sum_{x=0}^{n} \frac{Fi(x)}{SumFi(x)}$$

Step 3: *if(i==0) Cluster_List[i]* $\leftarrow fx.itemset$

Step 4: *if(Cluster_List[i…n].equals(*$fx.itemset$*))*

Cluster_List[new] $\leftarrow fx.itemset$

Step 5 *: Cluster_List optimized with descending order*

Step 6 : *end fucntion*

mSVM classifier

Input: Normalized training dataset *Train_Data[]*, Normalized testing dataset *Test_Data[]*, defined threshold *qTh*

Output: Result set as output with *{Predicted_class, weight}*

Step 1: Read all test data from *Test_Data[]* using below function for validating to training rules, the data is normalized and transformed according to algorithms requirements

$$\text{test\_Feature(data)} = \sum_{m=1}^{n} (.\ \text{Attribute\_Set[A[m] … … . . A[n]} \leftarrow Test\_Data)$$

Step 2 : select the features from extracted attributes set of test_Feature(data) and generate feature map using below function.

$$\text{Test\_FeatureMap [t…….n]} = \sum_{x=1}^{n}(t) \leftarrow \text{test\_Feature(x)}$$

Test_FeatureMap [x] are the selected features in pooling layer. The convolutional layer extracts the features from input and passes to pooling layer and those selected features are stored in *Test_FeatureMap*

Step 3: Now read entire taring dataset to build the hidden layer for classification of entire test data in sense layer,

$$\text{train\_Feature(data)} = \sum_{m=1}^{n} (.\ \text{Attribute\_Set[A[m] … … . . A[n]} \leftarrow Train\_Data)$$

Step 4 : Generate the training map using below function from input dataset

$$\text{Train\_FeatureMap [t…….n]} = \sum_{x=1}^{n}(t) \leftarrow \text{train\_Feature(x)}$$

Train_FeatureMap[t] is the hidden layer map that generates feature vector for build the hidden layer. That evaluate the entire test instances with train data.

Step 5 : After generating the feature map we calculate similarity weight for all instances in dense layer between selected features in pooling layer

$$\text{Gen\_weight} = \text{CalcWeight (Test\_FeatureMap} \,||\, \sum_{i=1}^{n} \text{Train\_FeatureMap[i])}$$

Step 6 : Evaluate the current weight with desired threshold

$$\text{if(Gen\_weight} >= \text{qTh)}$$

Step 7 : $\text{Out\_List.add (trainF.class, weight)}$

Step 8 : Go to step 1 and continue when Test_Data == null

Step 9 : Return Out_List

The above describes a classification technique used after the dimensionality reduction in optimization techniques. In steps 1 and 2, the feature extraction selection process for testing data, while a similar task was done in steps 3 and 4 for training data. Step 5 demonstrates the generate an equal weight for test instance according to all training rules, and based on the achieved weight it predicts the final class label.

Results and Discussions

In this study, we have done a heterogeneous experimental setup to evaluate the proposed work. The open-source environment has been used with JDK 1.8 with NetBeans 8.0. The 2.7 GHz processor has been used with 8 GB RAM. All experiments are independently executed in a similar environment and demonstrated the entire obtained result in table 1.

Table 1: dimensionality reduction with various optimization techniques and clustering accuracy with K-means and SVM based proposed classification technique

Method	Input Dataset (MB) (Health Care)	After dimensionality reduction (MB)	K-means	SVM (Proposed)
GA	500 MB (VM Log Dataset)	450	81.80%	86.20%
ACO		430	83.00%	87.10%
PCO		400	85.40%	86.30%
ACO-GA		388	87.60%	90.70%
Proposed		**350**	**91.60%**	**93.40%**

The above Table 1 describes the dimensionality reduction results from various Optimization algorithms on the VM log dataset. After the processing of dimensionality reduction, validation has done with unsupervised clustering algorithms that provide accuracy for each by using K-means and proposed clustering method.

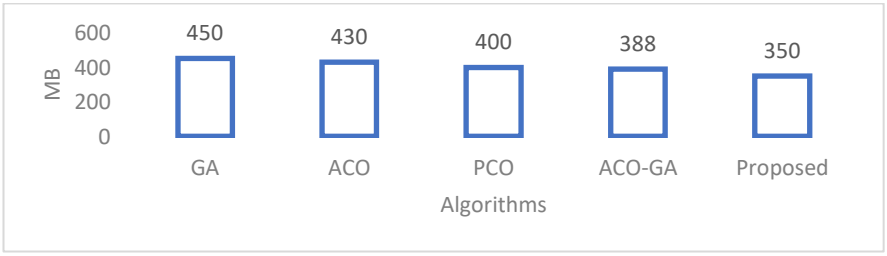

Figure 2 : Data optimization using various optimization techniques for dimensionality reduction (input data size-500 MB)

The figure 2 demonstrates dimensionality reduction of data after applying five different optimization algorithms. The proposed Q-learning based optimization technique reduces higher noisy data than other optimization techniques.

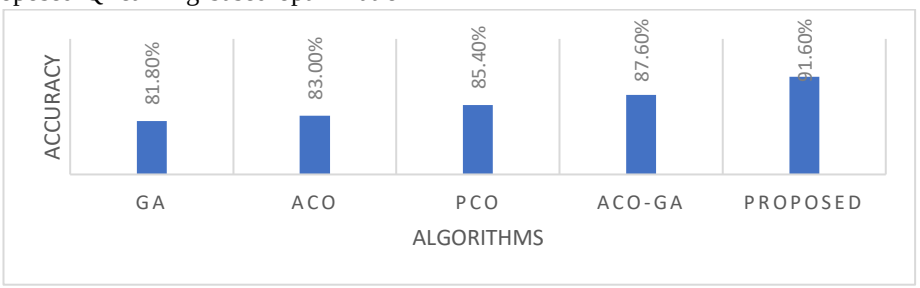

Figure 3 : classification accuracy of k-means with various optimization techniques

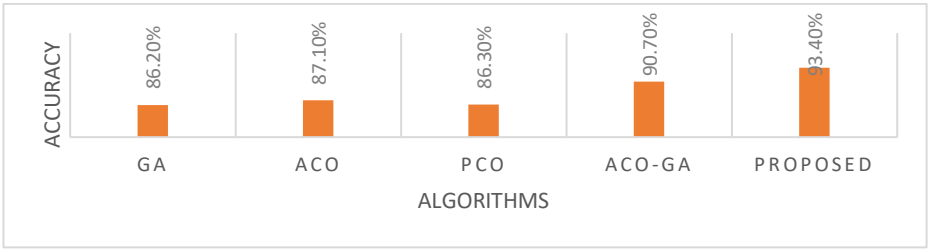

Figure 4: classification accuracy of proposed mSVM with various optimization techniques

The above Figure 3 and 4 demonstrates classification accuracy with K-means classification and proposed mSVM classifier. The K-means achieves 91.60% accuracy and mSVM obtains 93.40% using Q-learning based dimensionality techniques.

Conclusion

In this research, we proposed reinforcement-based clustering techniques for dimensionality reduction and cluster generation, while mSVM has been used for classification. We have demonstrated five different optimization techniques for generating optimized clusters and two different machine learning classifiers for evaluating the system's performance. The Q-Learning achieved higher results for dimensionality reduction while mSVM achieved 93.50% accuracy for classification. To evaluate the various deep learning algorithms on extensive unstructured text data will be the future direction of this system.

References

[1]. Mohamed Aymen Ben HajKacem, Chiheb-Eddine Ben N and Nadia Essoussi. “KP-S: A Spark-based Design of the K-Prototypes Clustering for Big Data”, 2017, ACS 14th International Conference on Computer Systems and Applications, IEEE

[2]. Ms. Tejaswini U. Mane. “Smart heart disease prediction system using Improved K-Means and ID3 on Big Data”, 2017, International Conference on Data Management, Analytics and Innovation (ICDMAI), IEEE

[3]. Sheela Gole and Bharat Tidke. “Frequent Itemset Mining for Big Data in social media using ClustBigFIM algorithm”, 2015, International Conference on Pervasive Computing (ICPC), IEEE

[4]. Jian Yin, Zhi-Fang Tan, Jiang-Tao Ren and Yi-Qun Chen. “An Efficient Clustering algorithm For Mixed Type Attributes In Large Dataset”, 2005, Proceedings of the Fourth International Conference on Machine Learning and Cybernetics, Guangzhou, IEEE

[5]. Ahmed Elgohary and Mohamed A. Ismail. “Efficient Data Clustering Over Peer-to-Peer Networks”, 2011, IEEE

[6]. Yu-Fang Zhang, Jia-Li Mao and Zhong-Yang Xiong. “An Efficient Clustering algorithm”, 2003, Proceedings of the Second International Conference on Machine Learning and Cybernetics, Xi’an, IEEE

[7]. Sheng-Yi Jiang and W-Ming Xu. “An Efficient Clustering algorithm”, 2004, Proceedtags of the Third International Conference on Machine Learning and Cybernetics, Shanghai, IEEE

[8]. Sheng-Yi Jiang and Xia Li. “A Hybrid Clustering Algorithm”, 2009, Sixth International Conference on Fuzzy Systems and Knowledge Discovery, IEEE

[9]. Hui Zhang, Bin Pang, Ke Xie, and Hui Wu. “An Efficient Algorithm for Clustering Search Engine Results”, 2006, IEEE

[10]. Rasim Alguliyev, Ramiz Aliguliyev, Adil Bagirov and Rafael Karimov. “Batch Clustering Algorithm for Big Data Sets”, 2020, IEEE

[11]. Carlos Ordonez, Sikder Tahsin Al-Amin and Ladjel Bellatreche. “An ER-Flow Diagram for Big Data”, 2020, International Conference on Big Data (Big Data), IEEE

[12]. jian Yin and Dongfang Zhao. “Data Confidentiality Challenges in Big Data Applications”, 2015, International Conference on Big Data (Big Data), IEEE

[13]. Luo Xiaofeng and Luo Jing. “Research on Big Data Reference Architecture Model”, 2020, International Conference on Artificial Intelligence and Big Data, IEEE

[14]. Ikbal Taleb and Mohamed Adel Serhani. “Big Data Pre-Processing: Closing the Data Quality Enforcement Loop”, 2017, 6th International Congress on Big Data, IEEE

[15]. Sarma, C. A. ., S. . Inthiyaz, B. T. P. . Madhav, and P. S. . Lakshmi. “An Inter Digital- Poison Ivy Leaf Shaped Filtenna With Multiple Defects in Ground for S-Band Bandwidth Applications”. International Journal on Recent and Innovation Trends in Computing and Communication, vol. 10, no. 8, Aug. 2022, pp. 55-66, doi:10.17762/ijritcc.v10i8.5668.

[16]. David Becker and Bill McMullen. “Big Data, Big Data Quality Problem”, 2015, International Conference on Big Data (Big Data), IEEE

[17]. Ernesto Damiani. “Toward Big Data Risk Analysis”, 2015, International Conference on Big Data (Big Data), IEEE

[18]. Bhagyashri S. Gandhi and Leena A. Deshpande. “The Survey on Approaches to Efficient Clustering and Classification Analysis of Big Data”, 2020, IEEE

[19]. Disha D N, Sowmya B J, Chetan and Dr. Seema S. “An Efficient Framework of Data Mining and its Analytics on Massive Streams of Big Data Repositories”, 2016, IEEE

[20]. Zakaria Gheid and Yacine Challal. “Efficient and Privacy-Preserving k-means clustering For Big Data Mining”, 2016, IEEE TrustCom/BigDataSE/ISPA

[21]. Maryam Abdullah, Fawaz S. Al-Anzi and Salah Al-Sharhan. “Efficient Fuzzy Techniques for Medical Data Clustering”, 2017, 9th IEEE-GCC Conference and Exhibition (GCCCE), IEEE

[22]. Doaa.Sayed, Sherine.Rady and Mostafa.Aref. “Enhancing CluStream Algorithm for Clustering Big Data Streaming over Sliding Window”, 2020, IEEE

[23]. Louis Y. Y Lu and John S. Liu. “The major research themes of big data literature:”, 2016, International Conference on Computer and Information Technology, IEEE

[24]. Xinxin Huang and Shu Gong. “Analysis of Big-Data Based Data Mining Engine”, 2017, 13th International Conference on Computational Intelligence and Security, IEEE

[25]. Galina Chernyshova, Gennady Smorodin and Alexey Ovchinnikov. “Technique of Cluster Validity for Text Mining”, 2016, IEEE

[26]. M. Omair Shafiq. “Event Segmentation using MapReduce based Big Data Clustering”, 2016, International Conference on Big Data (Big Data), IEEE

[27]. Dr. Anu Saini, Jagrit Minocha, Jaypriya Ubriani and Dhruv Sharma. “New Approach for Clustering of Big Data: DisK-Means”, 2016, International Conference on Computing, Communication and Automation (ICCCA), IEEE

[28]. Seema Maitrey and C.K. Jha. “Handling Big Data Efficiently by using Map Reduce Technique”, 2015, International Conference on Computational Intelligence & Communication Technology, IEEE

[29]. Dajung Lee, Alric Althoff, Dustin Richmond and Ryan Kastner. “A Streaming Clustering Approach Using a Heterogeneous System for Big Data Analysis”, 2017, IEEE

[30]. Suyash Mishra and Dr Anuranjan Misra. “Structured and Unstructured Big Data Analytics”, 2017, International Conference on Current Trends in Computer, Electrical, Electronics and Communication (ICCTCEEC), IEEE

[31]. Ni Bin. “Research on Methods and Techniques for IoT Big Data Cluster Analysis”, 2018, International Conference on Information Systems and Computer Aided Education (ICISCAE), IEEE

[32]. S. Dhanasekaran, R. Sundarrajan, B. S. Murugan, S. Kalaivani and V. Vasudevan. “Enhanced Map Reduce Techniques for Big Data Analytics based on K-Means Clustering”, 2019, IEEE

[33]. Agarwal, D. A. . (2022). Advancing Privacy and Security of Internet of Things to Find Integrated Solutions. International Journal on Future Revolution in Computer Science &Amp; Communication Engineering, 8(2), 05–08. https://doi.org/10.17762/ijfrcsce.v8i2.2067

[34]. Ankita Saldhi, Abhinav Goel, Dipesh Yadav, Ankur Saldhi, Dhruv Saksena and S. Indu. “Big Data Analysis Using Hadoop Cluster”, 2014, IEEE

[35]. Saurabh Arora and Inderveer Chana. “A Survey of Clustering Techniques for Big Data Analysis”, 2014, IEEE

[36]. G. Anuradha and Bidisha Roy. “Suggested Techniques for Clustering and Mining of Data Streams”, 2014, International Conference on Circuits, Systems, Communication and Information Technology Applications (CSCITA), IEEE

[37]. Doaa.Sayed, Sherine.Rady and Mostafa.Aref. “Enhancing CluStream Algorithm for Clustering Big Data Streaming over Sliding Window”, 2020, IEEE

[38]. Charalampos Chelmis, Jahanvi Kolte and Viktor K. Prasanna. “Big Data Analytics for Demand Response: Clustering Over Space and Time”, 2015, International Conference on Big Data (Big Data), IEEE

[39]. Ishwank Singh, A Sai Sabitha and Abhay Bansal. “Student Perfoemance Analysis Using Clustering Algorithm”, 2016, IEEE

[40]. Giannis Spiliopoulos, Konstantinos Chatzikokolakis, Dimitrios Zissis, Evmorfia Biliri, Dimitrios Papaspyros, Giannis Tsapelas and Spyros Mouzakitis. “Knowledge extraction from maritime spatiotemporal data: An evaluation of clustering algorithms on Big Data”, 2017, International Conference on Big Data (BIGDATA), IEEE

[41]. Fadia Alaeddin, Ala’ Khalifeh and Khalid A. Darabkh. “An Overview on Big Data Mining Using Evolutionary Techniques”, 2020, International Conference on Innovation and Intelligence for Informatics, Computing and Technologies (3ICT), IEEE

[42]. Linda R. Musser. (2020). Older Engineering Books are Open Educational Resources. Journal of Online Engineering Education, 11(2), 08–10. Retrieved from http://onlineengineeringeducation.com/index.php/joee/article/view/41

[43]. Jungkyu Han and Min Luo. “Bootstrapping K-means for Big data analysis”, 2014, International Conference on Big Data, IEEE

[44]. Neha Bharill, Aruna Tiwari and Aayushi Malviya. “Fuzzy Based Scalable Clustering Algorithms for Handling Big data using Apache Spark”, 2016, IEEE

[45]. Sunil Kumar and Maninder Singh. “A Novel Clustering Technique for Efficient Clustering of Big Data in Hadoop Ecosystem”, 2019, IEEE

[46]. Kumar, S., Gornale, S. S., Siddalingappa, R., & Mane, A. (2022). Gender Classification Based on Online Signature Features using Machine Learning Techniques. International Journal of Intelligent Systems and Applications in Engineering, 10(2), 260–268. Retrieved from https://ijisae.org/index.php/IJISAE/article/view/2020

[47]. Garima, Hina Gulati and P.K.Singh. “Clustering Techniques in Data Mining: A Comparison”, 2019, IEEE