International Journal of Advanced Computer Technology

ISSN: 2319-7900 www.ijact.org Volume 12, Issue 1, February 2023

IMPROVEMENT OF DATA ANALYSIS BASED ON K-MEANS ALGORITHM AND AKMCA

Zeeshan Ali Khan, Manjari Singh, Rajesh Boghey

Computer Science Engineering Department,

Technocrats Institute of Technology Excellence, MP, India

Abstract: Data analysis is improved using the k-means algorithm and AKMCA. Data mining aims to extract information from a large data set and transform it into a functional structure. Exploratory data analysis and data mining applications rely heavily on clustering. Clustering is grouping a set of objects so that those in the same group (called a cluster) are more similar to those in other groups (clusters). There are various types of cluster models, such as connectivity models, distribution models, centroid models, and density models. Clustering is a technique in data mining in which the set of objects is classified as clusters. Clustering is the most important aspect of data mining. The algorithm makes use of the density number concept. The high-density number point set is extracted from the original data set as a new training set, and the point in the high-density number point set is chosen as the initial cluster centre point. The basic clustering technique and the most widely used algorithm is K-means clustering.

K-Means, a partition-based clustering algorithm, is widely used in many fields due to its efficiency and simplicity. However, it is well known that the K-Means algorithm can produce suboptimal results depending on the initial cluster centre chosen. It is also referred to as Looking for the nearest neighbours. It simply divides the datasets into a specified number of clusters. Numerous efforts have been made to improve the K-means clustering algorithm's performance. Advanced k-mean clustering algorithm (AKMCA) is used in data analysis to obtain useful knowledge of various optimisation and classification problems that can be used for processing massive amounts of raw and unstructured data. Knowledge discovery provides the tools needed to automate the entire data analysis and error reduction process, where their efficacy is investigated using experimental analysis of various datasets. The detailed experimental analysis and a comparison of proposed work with existing k-means clustering algorithms. Furthermore, it provides a clear and comprehensive understanding of the k-means algorithm and its various research directions.

Keywords: Data Mining, Supervised Learning, Unsupervised Learning, Clustering, K-means Clustering, Smart Data Analysis, VSM, Error Rate.

I. INTRODUCTION

Data mining is the extraction or mining of knowledge from large amounts of data. It is also defined as locating hidden data in a database. It is a technique that converts raw data into user-understandable information and is primarily used for discovering unknown patterns. It is increasingly being used in science and technology to extract vast amounts of data. Classification is the separation of data based on similarities in their characteristics. Some classification methods include the Nave Bayes Classifier, Decision Tree, Neural Networks, and Support Vector Machine [1]. Data mining is a multi-disciplinary subfield of computer science technology that is fundamentally a computing process for discovering patterns in large data sets. It is an essential process that includes intelligent methods for retrieving data patterns. Data mining is the practice of analysing massive existing datasets to generate new information, also known as the process of knowledge discovery from databases. In most applications, the data or knowledge gained from data mining is potentially new and very profitable; however, the extracted data is typically reserved for later use.

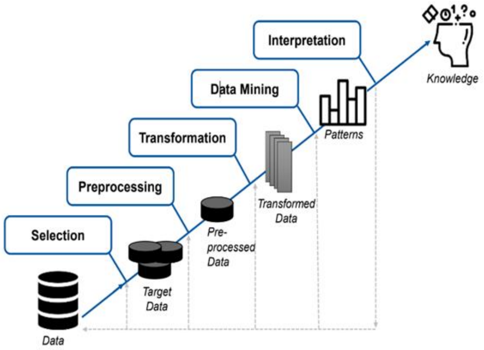

Fig1. Data Mining as a step in the process of KDD

It uses one or more software programmers to examine large amounts of data to extract information. In this world where business tactics are everywhere and every minute, data mining techniques and methodologies evolve as efficient methodologies for analysing raw data and converting it into insight that aids in making important business decisions. Specifically, data mining techniques dig deep into the materials to help us understand the various aspects. For example, the massive amount of information from the internet providing details on various fields could be a collection of raw data containing details. If properly sorted out, the same data can become knowledge or insight, leading to a specific development, which can be done using data mining [2]. The collection of raw data from databases such as relational, data warehouse, advanced and information reserves, object-oriented and object-relational, transaction and spatial, heterogeneous and legacy, multimedia and streaming, text, text mining, and web mining are all part of the data mining process. The process involved in data mining to derive insights has earned them various names, such as knowledge discovery, information harvesting, pattern analysis, and knowledge extraction. These names help us understand the usefulness of the data, paving the way for constructive measures in the area involved. Figure 3 depicts the data mining methods used to extract useful information from large amounts of data [3].

1.1 Clustering in Machine Learning: It is an unsupervised learning method in which we draw references from datasets consisting of input data without labelled responses. Generally, it is used to find meaningful structure, explanatory underlying processes, generative features, and groupings inherent in a set of examples. Clustering is dividing the population or data points into some groups such that data points in the same groups are more similar to other data points in the same group and dissimilar to the data points in other groups. It is a collection of objects based on similarities and dissimilarities between them.

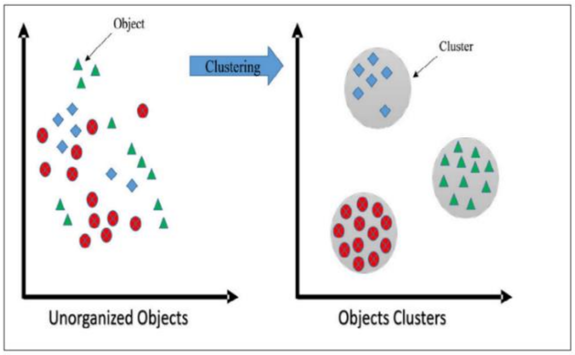

Fig2. Clustering in Machine Learning

II. RELATED WORK

Research on data mining in clustering has been done over the past few years. Some important K-mean clustering techniques are discussed here. Various clustering techniques have been proposed, and research has been done on unsupervised learning methods. Lanlan Zhang et al. [11] The D-K-means algorithm is proposed in this paper to address the shortcomings of the traditional K-means method. The algorithm makes use of the density number concept. The high-density number point set is extracted from the original data set as a new training set, and the point in the high-density number point set is chosen as the initial cluster centre point. The cluster centre points at high-density points are then updated using the geometric centre point method until convergence conditions are met. Experiments show that the method can effectively avoid the K-means clustering algorithm's local optimal situation.

On the other hand, the number of iterations in the clustering process is reduced, and clustering stability and accuracy are improved. Du et al. [12] Because the initial clustering centres of the traditional K-means algorithm are generated randomly from a data set, the clustering effect is not very stable. In order to solve this problem, this paper proposes a kind of optimal selection of the initial clustering centre of the K-means algorithm based on density, in which, after calculating the local density of each data point and the minimum distance between that point and any other point with higher local density, K points with higher local density are chosen as the initial clustering centres. Using the UCI standard database for contrast experiments demonstrated that the improved K-means algorithm can eliminate reliance on the initial clustering centre and has higher accuracy and stability than the traditional algorithm.

Ioniţa et al. [13] The healthcare sector generates massive data on patients and their illnesses, health insurance plans, medication and treatment schedules for various diseases, medical services, and so on. There is a growing demand in the healthcare community to transform existing quantities of healthcare data into value-added data by discovering unknown patterns and relationships between these data and using them in decision-making processes, regardless of whether they refer to management, planning, or treatments. Data mining is discovering knowledge and techniques such as classification and regression trees, logistic regression, and neural networks capable of predicting a patient's health status by considering various medical parameters (also known as attributes) and demographic parameters. This paper presents a case study on classifying patients with thyroid dysfunctions into three classes (i.e., 1 - hypothyroidism, 2 - hyperthyroidism, and 3 - normal) using data mining algorithms. It discusses potential methods to improve the accuracy of the considered classification models. Joseph et al. [14] The intelligent computing system, a collection of connected devices cooperating to achieve a specific goal, combines artificial intelligence and computational intelligence in various applications. The paper presents a survey on data mining algorithms and techniques that could be used with intelligent computing systems, presenting a basic concept of data mining along with the prominent algorithms of data mining and the classification of its techniques. The survey concludes with the challenges included in the overview and future improvements in the research that analyses data mining techniques in t. Parvathi et al. [15] Data mining is the process of extracting hidden information from a large set of databases. It can assist researchers in gaining novel and deep insights into large biomedical datasets. Data mining can uncover new biomedical and healthcare knowledge for clinical decision-making. This review first introduces data mining in general (e.g., definition, data mining tasks, data mining applications) and then provides a summary of various algorithms used for classification, clustering, and association. There is a discussion to enable disease diagnosis and prognosis, as well as the discovery of hidden biomedical and healthcare patterns from related databases, and a discussion of the use of data mining to discover such relationships as those between health conditions and disease. Diseases have relationships. It discusses the tool that can be used for data processing and classification and the benefits of WEKA. Moertini et al. [16] implemented a technique that improves the parallel implementation of the traditional K-means algorithm using MapReduce. The enhancement techniques are as follows: 1) data preprocessing, which includes attribute selection, cleaning, and transformation in addition to the clustering process. 2) Reducing the number of iterations by computing centroids initialisation in the map function. 3) Creating clustering patterns and clusters of quality measures. When the technique was tested on higher-spec computers, they discovered that the MapReduce-based K-means algorithm scaled better when run on two computers. Zhao et al. [17] The MapReduce model, which Hadoop implemented, used the K-means algorithm to build the clustering process applicable to big data analysis. The proposed algorithm's performance was evaluated using speedup, scale-up, and size-up criteria. According to the evaluation metrics, the presented algorithm performed admirably. Due to its efficiency and simplicity, Wei et al. [18] K Means is a partition-based clustering algorithm widely used in many fields. However, it is well known that the K-Means algorithm can produce suboptimal results depending on the initial cluster centres chosen. We propose a projection-based K-Means initialisation algorithm in this paper. The proposed algorithm first employs the traditional Gaussian kernel density estimation method to find the highly dense data areas in one dimension. The projection step then involves iteratively using density estimation from lower variance dimensions to higher variance dimensions until all dimensions are computed. Experiments on real-world datasets show that our method can achieve comparable results to conventional methods with fewer computation tasks. Haobin et al. [19] address the shortcomings of the traditional data classification method. This paper proposes a method for classifying data that employs the genetic and K-means algorithms. To improve the effectiveness of data analysis, and because the initial cluster centre easily influences the classical K-means algorithm with random selection, this paper improves the K-means algorithm by optimising the initial cluster centre. The sorted neighbourhood method (SNM) is used to preprocess the data in this paper, and then the K-means algorithm is used to cluster the data. This paper optimises the initial cluster centre and unifies the genetic algorithm for data dimensionality reduction to improve the accuracy of the K-means algorithm. The results of the experiments show that the proposed method outperforms the traditional data classification method in terms of classification accuracy. Dileep Kumar et al. [20] Database mining is the process of mining for implicit, previously unidentified, and potentially vital information from massive databases using efficient knowledge discovery techniques. The privacy and security of user information have become significant public policy concerns. These concerns have piqued the interest of both public and government legislators and controllers, privacy advocates, and the media. We focus on key online privacy and security issues and concerns in this paper, as well as the role of self-regulation and the user in privacy and security protections, data protection laws, regulatory trends, and the outlook for privacy and security legislation. Naturally, such a process opens up new assumption dimensions, identifies new invasion patterns, and raises new data security issues in information technology advancements in recent years.

III. PROPOSED METHODOLOGY

Data Mining refers to discovering "interesting" patterns in large data collections. The data industry has a large amount of data available. This knowledge is useless until it is regenerated into useful information. It's necessary to research this large quantity of data and extract helpful information from it. Data mining describes the abstract goals of what must be done. It relies on a wide range of different techniques to achieve them, such as artificial neural networks, cluster analysis, and different processes such as data cleaning, data integration, information transformation, data mining, pattern analysis, and knowledge extraction. Presentation

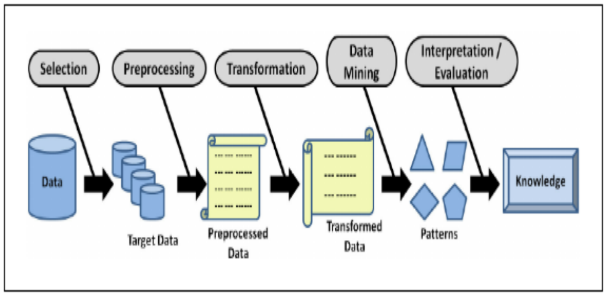

Fig3. block diagram of information detection Process

When these processes overcome data error problems, they may be able to use this data in various applications such as error detection in large data sets, market dataset analysis research, production control, science exploration, and so on. Clustering can also be defined as an information reduction tool, as it generates larger and more manageable subgroups than individual data points. Clustering is justified as a method for grouping a wide range of knowledge into significant teams or clusters based on data similarity types. Clusters are teams with similar knowledge based on common options but dissimilar to knowledge in other clusters. Knowledge objects created by cluster analysis teams are primarily based solely on data found in the knowledge of various groups. The knowledge objects of cluster analysis teams are primarily based on data found in the knowledge that describes the objects and their relationships. AKMCA-based healthcare dataset enhancement. An analysis of a dataset for infectious disease detection and prevention using data mining techniques. Improved performance of k-means clustering for communicable disease dataset classification and proposed algorithm (AKMCA). Thyroid Dataset, E-coli Dataset, Iris dataset, and breast cancer dataset are improving classification by using intelligent data analysis modal (vector space) copy data reduction. K-means clustering determines the number of clusters, but copy data has a high error rate, and we also know that the result is suboptimal (error data). The healthcare dataset is loaded in the initialisation process and then sets any data point value in the dataset. Clustering is performed using K-means, and the analysis of error values and the proposed algorithm (AKMCA) are used to reduce dataset error values. The proposed algorithm (AKMCA) generates initial clusters. It eliminates copy data to improve cluster accuracy and reduce time and error but requires more iterations than the existing K-means clustering method.

IV. SIMULATION TOOL

This paper implements MATLAB to fast execution and computation performed on the input cancer dataset. It uses its JIT (just in time) compilation technology to provide execution speeds that rival traditional programming languages. It can also use multicore and multiprocessor computers to provide multi-threaded linear algebra and numerical functions. These functions automatically execute on multiple computational threads in a single MATLAB to execute faster on multicore computers. This thesis performed all enhanced efficient data retrieval results in MATLAB.MATLAB is the high-level language and interactive environment used by millions of engineers and scientists worldwide. It enables users to explore and visualise ideas while collaborating across disciplines through signal and image processing, communication, and result computation. MATLAB provides tools for collecting, analysing, and visualising data, allowing you to gain insight into your data in a fraction of the time it would take using spreadsheets or ancient programming languages.

Table 1 Results analysis of the breast cancer dataset

Algo	SRV	Time	ER (%)	ITR
KMAD	0.4743	1.2188	4.557	4
AKMCA	0.4743	1.4063	1.747	5
KMAD	0.9525	0.1094	2.6552	3
AKMCA	0.9525	0.0313	0.0087	4
SRV=Set Random Values, ER=error rate, ITR=Iteration				

It can also document and share the results via plots and reports and reveal MATLAB code. MATLAB (matrix laboratory) is a fourth-generation artificial language and multi-paradigm numerical computing scenario. It was created through mathematical work; MATLAB allows for matrix strategy, function and data plotting, algorithm implementation, and the creation of user interfaces with programmers. MATLAB is intended mainly for mathematical computing; an optional toolbox uses the MuPAD symbolic engine, allowing access to symbolic computing capabilities.

V. RESULT ANALYSIS

a) Comparison between KMAD and AKMCA Based on breastcancer_dataset in case1. Results analysis performed on breast cancer dataset KMAD and AKMCA in which KMAD are error is more and time also more but iteration min and AKMCA error less and average time but iteration more.

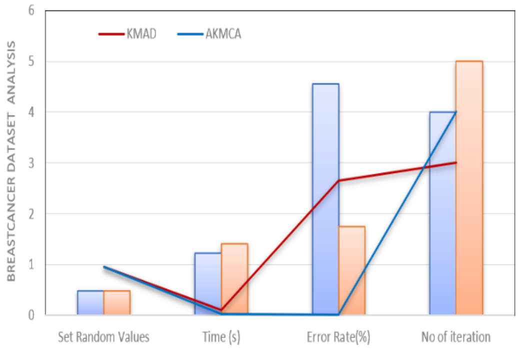

Fig4. breast cancer dataset analysis between KMAD and AKMCA

Show results in figure 4 on the breast cancer dataset between KMAD and AKMCA. The KMAD dataset contains more errors and time but iteration minimum, and AKMCA errors less and average time but iteration more.

VI. CONCLUSION

Data analysis is improved using the k-means algorithm and AKMCA. The proposed algorithm's overall goal is to extract information from a large data set and transform it into a usable format for further use. Clustering is a critical task in data analysis and mining applications. Clustering is grouping objects so that objects in the same group are more similar. Clustering algorithms share some common issues that must be addressed to succeed. Some issues are so widespread that they are not even unique to unsupervised learning and can be considered part of a broader data mining framework. Other issues are addressed in specific algorithms; they presented the popular k-means algorithm and issues with initialisation and the inability to handle data with mixed features. Unlike previous studies, this paper includes a critical analysis of existing methods and an experimental analysis of machine learning datasets to demonstrate the performance of various k-means variants. The experimental analysis revealed no universal solution to the k-means algorithm's problems; each of the algorithm's existing variants is either application-specific or data-specific. Our future research will concentrate on developing a robust k-means algorithm capable of addressing both problems concurrently. After projecting this step for all dimensions, they can construct a good but suboptimal initial centre for the K-Means algorithm. Compared to other proposed methods, our proposed (AKMCA) algorithm performs fewer computation tasks but with comparable accuracy. The proposed algorithm improves information optimisation and error minimisation, increases iteration and average time, and obtains optimal solutions. Data analysis is improved using the k-means algorithm and AKMCA. Overall result analysis summary Discover our proposed algorithm's error rate depreciation compared to the existing method and our proposed algorithm's average time but more iterations compared to the existing method. We used four datasets for analysis: breast cancer dataset, E-coil dataset, thyroid dataset, and iris dataset. Our proposed algorithm yields optimal and dependable results.

REFERENCES

[1]. Bandyopadhyay, Seema, and Edward J. Coyle. "An energy-efficient hierarchical clustering algorithm for wireless sensor networks." In IEEE Infocom 2003. Twenty-second Annual Joint Conference of the IEEE Computer and Communications Societies, vol. 3, pp. 1713-1723. IEEE, 2003.

[2]. A. Kumar, R. Sinha, V. Bhattacharjee, D. S. Verma, S. Singh, "modelling using K-means clustering algorithm", IEEE 2012, 1st international conference on recent advances in information technology (RAIT).

[3]. Bruno Fernandez Chimieski, Rubem Dutra RibeiroFagundes, "Association and Classification Data Mining Algorithms Comparison over Medical Datasets", J. Health Inform. Abril-Junho; 5(2): 44-5, 2013.

[4]. D. Arthur, S. Vassilvitskii, "k-means++: The advantages of careful seeding", Proceedings of the 18th annual ACM-SIAM symposium on discrete algorithms, pp. 1027–1035, 2007.

[5]. Ester, Martin, Hans-Peter Kriegel, Jörg Sander, and Xiaowei Xu. "A density-based algorithm for discovering clusters in large spatial databases with noise." In KDD, vol. 96, no. 34, pp. 226-231. 1996.

[6]. Joachims, Thorsten. "Text categorisation with support vector machines: Learning with many relevant features." In European conference on machine learning, pp. 137-142. Springer, Berlin, Heidelberg, 1998.

[7]. Junatao Wang, XiaolongSu, "An Improved K-means Clustering Algorithm, Communication Software and Networks (ICCSN), IEEE 3rd International Conference on 27 May, (pp. 44-46), 2011.

[8]. Kiri Wagstaff, Claire Cardie, Seth Rogers, Stefan Schroedl, "Constrained K-means Clustering with Background Knowledge", ICML Proceedings of the Eighteenth International Conference on Machine Learning, pp. 577–584, 2001.

[9]. K. A. Abdul Nazeer, M. P. Sebastian, "Improving the Accuracy and Efficiency of the k-means Clustering Algorithm, Proceedings of the World Congress on Engineering 2009 Vol I WCE 2009, July 1 - 3, London, U.K., 2009.

[10]. Kanungo, Tapas, David M. Mount, Nathan S. Netanyahu, Christine D. Piatko, Ruth Silverman, and Angela Y. Wu. "An efficient k-means clustering algorithm: Analysis and implementation." IEEE Transactions on Pattern Analysis & Machine Intelligence 7: 881-892, 2002.

[11]. Zhang, Lanlan, Jinshuai Qu, Minghu Gao, and Meina Zhao. "Improvement of K-means algorithm based on density." In 2019 IEEE 8th Joint International Information Technology and Artificial Intelligence Conference (ITAIC), pp. 1070-1073. IEEE, 2019.

[12]. Du, Xin, Ning Xu, Cailan Zhou, and Shihui Xiao. "A density-based method for selecting the initial clustering centres of K-means algorithm." In 2017 IEEE 2nd Advanced Information Technology, Electronic and Automation Control Conference (IAEAC), pp. 2509-2512. IEEE, 2017.

[13]. Ioniţa, Irina, and Liviu Ioniţa. "Applying data mining techniques in healthcare." Stud Inform Control 25, no. 3, 385-94, 2016.

[14]. Joseph, S. Iwin Thanakumar, and Iwin Thanakumar. "Survey of data mining algorithms for the intelligent computing system." Journal of trends in Computer Science and Smart technology (TCSST) 1, no. 01 (2019): 14-24.

[15]. Parvathi, I., and Siddharth Rautaray. "Survey on data mining techniques for diagnosing diseases in the medical domain." International Journal of Computer Science and Information Technologies 5, no. 1 (2014): 838-846.

[16]. Moertini V, Venica L . Enhancing Parallel k-Means Using Map Reduce for Discovering Knowledge from Big Data. IEEE Int Conf Cloud Computing Big Data Anal 81–87, 2016.

[17]. Zhao W, Ma H, He Q. Parallel K -Means Clustering Based on MapReduce. CloudCom 674–679, 2009.

[18]. Du, Wei, Hu Lin, Jianwei Sun, Bo Yu, and Haibo Yang. "A new projection-based K-Means initialisation algorithm." In 2016 IEEE Chinese Guidance, Navigation and Control Conference (CGNCC), pp. 2341-2345. IEEE, 2016.

[19]. Shi, Haobin, and Meng Xu. "A data classification method using genetic algorithm and K-means algorithm with optimising initial cluster centre." In 2018 IEEE International Conference on Computer and Communication Engineering Technology (CCET), pp. 224-228. IEEE, 2018.

[20]. Singh, Dileep Kumar, and Vishnu Swaroop. "Data security and privacy in data mining: research issues & preparation." International Journal of Computer Trends and Technology 4, no. 2: 194-200, 2013.

[21]. Marty, Babu, G.P. and M.N., "Clustering with evolution strategies Pattern Recognition", 27, 2, 321-329, 1994.

[22]. Mitchell, Tom M. "Machine learning and data mining." Communications of the ACM 42.11, 1999.

[23]. Md Sohrab Mahmud, Md. Mostafizer Rahman, Md. Nasim Akhtar, "Improvement of K-means Clustering algorithm with better initial centroids based on weighted average", IEEE 7th International Conference on Electrical and Computer Engineering, pp. 647-650, 2012.