INTERNATIONAL JOURNAL OF INNOVATIVE RESEARCH IN TECHNOLOGY & SCIENCE

ISSN: 2321-1156 **WWW.IJIRTS.ORG** **VOLUME IX ISSUE VI, NOVEMBER 2021**

Improvement Performance of Data Analysis Using Unsupervised K-Means Clustering and PGM

Arpana Kumari[1], Monika Raghuvanshi [2]

Department of CSE, Bhabha Engineering Research Institute, MP, Bhopal, India

[1]arpana7info@gmail.com, [2]monipriya21@gmail.com

Abstract: *Improved data analysis function using k definition cluster and unsupervised PGM. Data mining is a way to extract meaningful knowledge from these large databases. There are a variety of applications in this field, so the introduction of new ways of distributing data is thoroughly researched to improve performance. Data mining algorithms group the algorithms implemented on the machine, and those also used to make intelligent machine learning are called unsupervised machine learning algorithms and can perform critical functions through a k-definition compilation algorithm. Based on a well-defined particle swarm optimization algorithm, is often more error in data analysis. More complex issues can be tackled and solved as more information becomes available. The examination of patient data becomes more essential to find the patient's state of health and prevent and take other preventive measures. With the help of technology and automated machines, data can be analyzed more efficiently. Managing large amounts of data poses many data security concerns. Tests on actual data have shown that our technology will perform against different metrics with smaller data processing functions. Mining activities and mining techniques open up new opportunities for detecting the disease. Likewise, to provide effective treatment for the disease for three years, data mining can be used; in the end, a significant positive effect was obtained. The algorithm is tested with a UCI data set. The proposed molecular method is to improve the performance of data analysis and obtain improved data from the collection and analysis of errors implemented using MATLAB software, overcoming challenges of various mining applications and implementing unmanned drill shops. A whole cluster review is done with existing strategies and an easy to compare course so that it will be easy for someone to choose a specific approach that suits their working environment.*

Keywords: *Data Mining, Supervised Learning, Unsupervised Learning, Unsupervised Machine Learning Algorithms, Clustering Method, K-Means Clustering, Dataset, Data Analysis.*

I. INTRODUCTION

Data mining explores and analyses large data sets to discover meaningful patterns and rules. The key idea is to find an effective way to combine the computer's power to process the data with the human eye's ability to detect patterns. The objective of data mining is designed for and work best with large data sets. Data mining is a broader process called knowledge discovery from databases [1]. Data mining is a multi-step process, requires accessing and preparing data for mining the data, data mining algorithm, analyzing results and taking appropriate action. The data, which is accessed, can be stored in one or more operational databases. In data mining, the data can be mined, bypassing various processes. The data is mined using two learning approaches, i.e., supervised learning or unsupervised learning.

Supervised Learning: In supervised learning (often also called directed data mining), the variables under investigation can be split into two groups: explanatory variables and one (or more) dependent variables. The goal of the analysis is to specify a relationship between the dependent variable and explanatory variables; as it is done in regression analysis to proceed with directed data mining techniques, the values of the dependent variable must be known for a sufficiently large part of the dataset.

Unsupervised Learning: The desired result is not provided to the unsupervised model during the learning procedure. This method can only cluster the class input data based on their statistical properties. These models are for various types of clustering, k-means, distances and normalization, self-organizing maps. In unsupervised learning, all the variables are treated the same way, and there is no distinction between dependent and explanatory variables.

However, in contrast to the name undirected data mining, still, there is some target to achieve. This target might be as data reduction as general or more specific like clustering. The dividing line between unsupervised learning and supervised learning is the same that distinguishes discriminate analysis from cluster analysis. Supervised learning requires target variable should be well defined and that a sufficient number of its values are given. In unsupervised learning, typically, the target variable has only been recorded for too small several cases, or the target variable is unknown [2, 3].

Clustering algorithms have many categories like hierarchical-based algorithms, partition-based algorithms, density-based algorithms and grid-based algorithms. Partition-based clustering is centroid based which splits data points into k partitions, and each partition represents a cluster. Kmeans is a clustering algorithm that is used widely. This technique will be helpful in the extraction of useful information using clusters from massive Databases [4]. The overall purpose of data mining is to extract useful information from a massive set of data and convert it into a form that is

understandable for further use. For example, Data clustering is a data exploration technique that allows objects with similar characteristics to be grouped to further processing. Clustering assigns data points with similar properties to the same group and different data points to different groups—members within a cluster exhibit similar characteristics to the members of other clusters. Clustering is a technique that divides data objects into groups based on the information found in data that describes the objects and relationships among them, their feature values which can be used in many applications, such as knowledge discovery, vector quantization, pattern recognition, data mining, data dredging [5]. There are mainly two techniques for clustering: hierarchical clustering and partitioned clustering. Data are not partitioned into a particular cluster in a single step, but a series of partitions takes place in hierarchical clustering, which may run from a single cluster containing all objects to n clusters, each containing a single object. And each cluster can have sub-clusters so that it can be viewed as a tree, a node in the tree is a cluster, the root of the tree is the cluster containing all the objects, and each node, except the leaf nodes, is the union of its children. But in partitioned clustering, the algorithms typically determine all clusters at once, it divides the set of data objects into non-overlapping clusters, and each data object is in exactly one cluster [6].

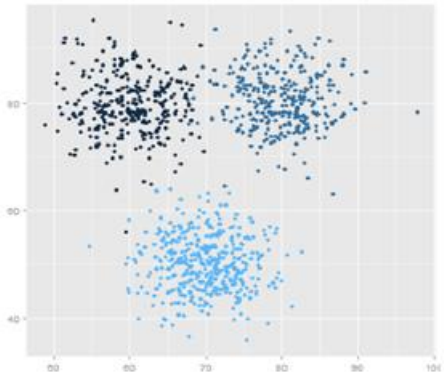

Fig.1 number of three clusters in clustering

II.RELATED WORK

L Bai et al. [7] In clustering, to find a better data clustering centre, make the algorithm convergence faster, and clustering results more accurate, a k-means clustering algorithm based on improved quantum particle swarm optimization algorithm is proposed. In this algorithm, the cluster centre is simulated as a particle. Cloning and mutation operations are used to increase the diversity and improve the global search ability of QPSO. A suitable and stable cluster centre is obtained. Finally, an effective clustering result is obtained. The algorithm is tested with the UCI dataset. The results show that the improved algorithm ensures the global convergence of the algorithm and obtains more accurate clustering results. It uses different breast cancer datasets from machine learning.

Shi et al. [8]. Aiming at the problems of the classical data classification method to propose a method using genetic algorithm and K-means algorithm to classify data. In order to improve the effectiveness of data analysis, considering that the classical K-means algorithm is easy to be influenced by the initial cluster centre with random selection, this paper improves the K-means algorithm by optimizing the initial cluster centre. This paper first uses the sorted neighbourhood method (SNM) to preprocess the data, and then the K-means algorithm is used to cluster data. In order to improve the accuracy of the K-means algorithm, this paper optimizes the initial cluster centre and unifies the genetic algorithm for the data dimensionality reduction. The experimental results show that the proposed method has higher classification accuracy than the classical data classification method.

Shafeeq et al. [9] present a changed K-means algorithm to spice up the cluster quality and fix the optimum cluster range as the user gives the input range of clusters (K) to the K-means algorithm. However, within the sensible state of affairs, it's tough to repair the number of clusters before. The strategy projected during this paper works for the cases, i.e., for a celebrated range of clusters, such as unknown clusters. The user has the flexibleness to fix the range of the number of clusters or input the minimum number of clusters needed. The algorithm computes the new cluster centres by incrementing the cluster counters in every iteration until it satisfies the cluster quality's validity. This algorithm can overcome this drawback by finding the optimum range of clusters on the run.

Kamaljit Kaur et al. [10] found that the K-Means algorithm has two significant limitations 1. Several distance calculations of each data point from all the centroids in each iteration. 2. The final clusters depend upon the selection of initial centroids. This work improves the k-Means clustering algorithm designed in MATLAB and the UCI machine learning repository datasets. The initial centroids were not selected randomly. By using a new approach, good clustering results were obtained. The new method of selecting the initial centroid is better than randomly selecting the initial centroids.

Junatao Wang et al. [11] propose an improved K-means algorithm using a noise data filter in this paper. This proposed algorithm overcomes the shortcomings of the traditional k-means clustering algorithm. The algorithm develops density-based detection methods based on characteristics of noise data where the discovery and processing steps of the noise data are added to the original algorithm. By preprocessing the data to exclude these noise data before clustering data sets, the cluster

cohesion of the clustering results is improved significantly, and the impact of noise data on the K-means algorithm is decreased effectively, and the clustering results are more accurate

Canlas et al. [12] The successful application of data mining in apparent fields like e-business, marketing, and retail has led to its popularity in knowledge discovery in databases (KDD) in other industries and sectors. Among these sectors that are just discovering data mining are medicine and public health. This research paper provides a survey of the current techniques of KDD, using data mining tools for healthcare and public health. It also discusses critical issues and challenges associated with data mining and healthcare in general. The research found a growing number of data mining applications, including analysis of health care centres for better health policy-making, detection of disease outbreaks and preventable hospital deaths, and detection of fraudulent insurance claims.

K. A. Abdul Nazeer et al. [13] propose a k-means algorithm that produces different clusters for different sets of values of initial centroids. The final cluster quality in the algorithm depends on the selection of initial centroids. Two phases include the original k means algorithm: first for determining initial centroids and second for assigning data points to the nearest clusters and then recalculating the clustering mean.

H. Altay Guvenir et al. [14] has planned a brand-new classification formula VFI5 and applied the drawback of medical diagnosis of erythematic squalors. Several authors have used the medical speciality dataset from UCI (the University of CA at Irvine), ranging from his work wherever he applied his new advanced formula VFI5. It represents an idea description by a group of feature intervals. The classification of a brand-new instance is predicated on a vote among the classification created by the values of every feature one by one. All training examples are processed quickly. The VFI5 formula constructs intervals for every feature from the training examples. For every interval, one price and, therefore, the votes of every category therein interval is maintained. Thus, an interval could represent many categories by sorting the vote for every category. This formula has obtained ninety-six—25% of classification accuracy.

Marty et al. [15] examine how the clustering technique can identify different information by considering various examples and seeing where the similarities and ranges agree. By examining one or more attributes or classes, you can group individual pieces of data to form a structured opinion. At a superficial level, clustering uses one or more attributes as your basis for identifying a cluster of correlating results. Clustering can work both ways. You can assume a cluster at a certain point and then use our identification criteria to see if you are correct.

II. SIMULATION TOOL

MATLAB (a shortened form of "Matrix Laboratory") is a restrictive multi-worldview programming language and numeric processing climate created by Math Works. MATLAB permits lattice controls, plotting capacities and information, execution of calculations, formation of UIs, and communicating with programs written in different dialects. Even though MATLAB is planned for numeric registering, a discretionary tool kit utilizes the MuPAD representative motor permitting admittance to emblematic figuring capacities. An extra bundle, Simulink, adds graphical multi-area reenactment and model-based plans for dynamic and installed systems. MATLAB apparatus could likewise be a bunch of hardware and execution for top execution numerical calculation and visual picture. It furnishes intuitive environmental elements with numerous sacred works for specialized calculation, illustrations and liveliness. The name MAT-LAB represents Matrix Laboratory. One in everything about component of MAT-LAB is its foundation autonomy. When you're in MATLAB, it doesn't make any difference which pc you're on for the preeminent half. In MATLAB, the M-records are the standard code text documents, with a .m expansion to the document name.

There are 2 documents of this record: script document and execution record. All most projects written in MATLAB are saved in M-records. Fig-documents are paired records with a. fig expansion which can be opened in MATLAB as figures. Such documents region unit made by saving a figure during this arrangement exploitation save or save as probability from the File menu or exploitation the save as order in order window-records are aggregated M-documents with a. p expansion which can be executed in MAT-LAB simple medical dataset analysis tool and function used.

IV.RESULT ANALYSIS

Data mining using k-means clustering-based cluster centre finds the minimum error of medical dataset analysis and the best possible solution. The infield of data mining and the number of challenges of the K-Mean clustering method based on unsupervised clustering algorithm improve dataset analysis performance, error minimizations using medical health care dataset analysis, and best solution. Experimentation Results Analysis using Different UCI Dataset : (a) Breast cancer wiscons in dataset Analysis:

The previous method (UKMCM) and the proposed method (PGM) using breast cancer wiscons dataset based result analysis here different random values set in

our dataset (breast cancer wiscons) like 0.7334, 0.1697, 0.521 and 0.4031 finally, overall compare UKMCM and PGM, UKMCM more error rate but PGM are less error rate and also not copy data. PGM best method of dataset analysis.

Table 1 Breast cancer Wiscon's dataset analysis

UCI Dataset	Set Random Values	Method	Time (In a sec)	Error Rate (%)
breast cancer Wisconsin dataset	0.7334	UKMCM	2.26201	4.1641
		PGM	3.77522	1.32556
	0.1697	UKMCM	1.99681	3.80586
		PGM	2.40242	1.08409
	0.521	UKMCM	4.58643	4.00499
		PGM	2.07481	1.21675
	0.4031	UKMCM	1.90321	3.92294
		PGM	2.27761	1.23968

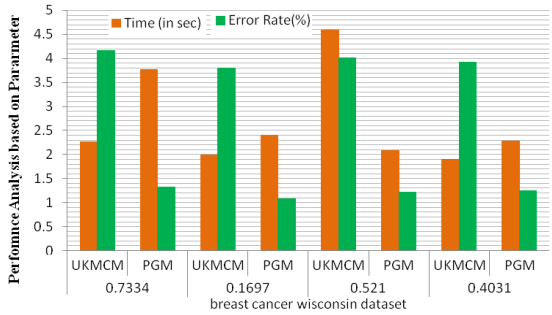

Fig 2 Breast cancer wiscons dataset-based compare analysis graph

Table 2 diabetes dataset analysis

Dataset	SRV	Method	Time	ER
Diabetes Dataset	0.6018	UKMCM	2.24641	4.57657
		PGM	4.05603	1.74665
	0.1028	UKMCM	2.74562	3.76561
		PGM	2.69882	1.04324
	0.5021	UKMCM	2.16841	4.57656
		PGM	3.13562	1.70782
	0.7081	UKMCM	6.55204	4.63243
		PGM	7.23845	1.91024
SRV= Set Random Values, **ER**= Error Rate				
Time in second, Error Rate in %				

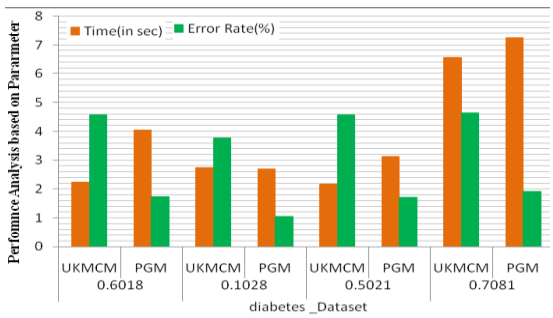

Fig 3 diabetes dataset-based compare analysis graph

(b) Diabetes Dataset in dataset Analysis: Previous method (UKMCM) and the proposed method (PGM) using diabetes dataset-based result analysis here different random values set in our dataset (diabetes dataset) like 0.6018, 0.1028, 0.5021and 0.7081 finally overall compare UKMCM and PGM, UKMCM more error rate but PGM are less error rate and also not copy data. PGM best method of dataset analysis.

Table 3 E. coli dataset analysis

Dataset	SRV	Method	Time	ER
E. coli _Dataset	0.1031	UKMCM	1.63801	3.78021
		PGM	2.40242	0.846778
	0.2027	UKMCM	4.35243	4.46619
		PGM	2.18401	1.29797
	0.5081	UKMCM	3.40082	5.1847
		PGM	1.99681	1.8197
	0.9102	UKMCM	0.327602	2.6552
		PGM	0.156001	0.0087105
SRV= Set Random Values, **ER**= Error Rate				
Time in second, Error Rate in %				

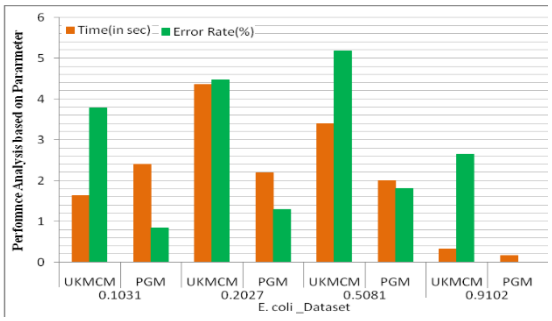

Fig 4 E. coli dataset-based compare analysis graph

Table 4 Yeast bacteria dataset analysis

Dataset	SRV	Method	Time	ER
Yeast bacteria dataset	0.9861	UKMCM	0.452403	2.6552
		PGM	0.218401	0.0087105
	0.0201	UKMCM	1.41961	3.28615
		PGM	0.889206	0.639659
	0.5098	UKMCM	2.09041	3.99423
		PGM	2.48042	1.21675
	0.7044	UKMCM	2.07481	4.01863
		PGM	3.75962	1.32556
SRV= Set Random Values, **ER**= Error Rate				
Time in second, Error Rate in %				

(c) E. coli dataset in dataset Analysis: Previous method (UKMCM) and the proposed method (PGM) using E. coli dataset-based result analysis here different random values set in our dataset (E. coli dataset) like 0.1031, 0.2027, 0.5081 and 0.910 finally overall, compare UKMCM and PGM, UKMCM more error rate but PGM are

less error rate and also not copy data. PGM best method of dataset analysis.

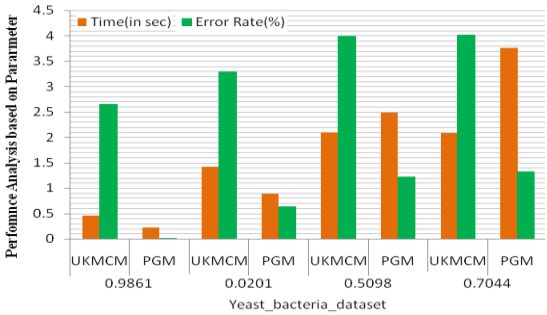

Fig 5 Yeast bacteria dataset-based compare analysis graph

(d) 4. Yeast bacteria dataset in dataset Analysis: The previous method (UKMCM) and the proposed method (PGM) use Yeast bacteria dataset based result analysis. Here are different random values set in our dataset (Yeast bacteria) like 0.9861, 0.0201, 0.5098and 0.7044.finally, overall, UKMCM and PGM UKMCM have more error rates, but PGM has fewer errors and does not copy data. PGM best method of dataset analysis.

V. CONCLUSION

Improve data analysis performance using unsupervised k-means clustering with PGM. Extracts data from a database and mining data to mine information from a vast data set and puts it in a functional form for other purposes and is used in business analysis. The collection is an essential function in data analysis and extraction applications. Proposed genetic clustering methods are essential for extensive data analysis process using unsupervised learning method using unsupervised learning approach and can be seen as part of the overall data management system. Many algorithms are explicitly designed to take a number of these problems, and k-mean focuses on those problems that can be solved independently in future research. Medical data management will facilitate the planning of specific diagnostic and decision-making strategies. The collection is how large databases break down into smaller databases called clusters. Some algorithms work well to collect data that can be divided into clusters, it uses different sets of breast cancer from machine learning, and the algorithm is tested with the set of UCI data. The K-Measurement algorithm offers different advantages and disadvantages in different types of K-Means applications. The analysis of the medical data set will design strategies to identify and decide the level of activation, focusing on the use of unsupervised K-Means based on machine learning for distribution. Improved data analysis performance using unsupervised PGM based on minimum error rate statistical value of medical dataset analysis.

REFERENCES

[1]. Xia fan, Chen, and Wang Feng in. "Application of data mining on enterprise human resource performance management." In 2010 3rd International Conference on Information Management, Innovation Management and Industrial Engineering, vol. 2, pp. 151-153. IEEE, 2010.

[2]. Jing, Han. "Application of fuzzy data mining algorithm in human resource performance evaluation." In 2009 International Forum on Computer Science-Technology and Applications, vol. 1, pp. 343-346. IEEE, 2009

[3]. M. Emre Celebi, H. A. Kingravi, P. A. Vela, "A Comparative Study of, Efficient Initialization Methods for the K-Means Clustering Algorithm", Expert Systems with Applications, pp. 200-210, vol.40, 2013.

[4]. Lu, J. F., Tang, J. B., Tang, Z. M., & Yang, J. Y, Hierarchical initialization approach for k-means clustering. Pattern Recognition Letters, 29(6), 787–795, 2008.

[5]. T. Zhang and Y. Bo, ``Density-based multiscale analysis for clustering in strong noise settings with varying densities,'' IEEE Access, vol. 6, pp. 25861_25873, 2018.

[6]. Redmond, S. J., & Heneghan, C., A method for initializing the k-means clustering algorithm using kd-trees, Pattern Recognition Letters, 28(8), 965–973, 2009.

[7]. Bai, Lili, Zerui Song, Haijie Bao, and Jingqing Jiang. "K-means Clustering Based on Improved Quantum Particle Swarm Optimization Algorithm." In 2021 13th International Conference on Advanced Computational Intelligence (ICACI), pp. 140-145. IEEE, 2021.

[8]. Shi, Haobin, and Meng Xu. "A Data Classification Method Using Genetic Algorithm and K-Means Algorithm with Optimizing Initial Cluster Center." 2018 IEEE International Conference on Computer and Communication Engineering Technology (CCET). IEEE, 2018

[9]. Shafeeq, A., Hareesha, K., Dynamic Clustering of Data with Modified K-Means Algorithm, International Conference on Information and Computer Networks, vol. 27,2012.

[10]. Kamaljit Kaur, Dr Dalvinder Singh Dhaliwal, Dr Ravinder Kumar Vohra," Statistically Refining the Initial Points for K-Means Clustering Algorithm ", International Journal of Advanced Research in Computer Engineering & Technology (IJARCET) Volume 2, Issue 11, November 2013.

[11]. Junatao Wang, XiaolongSu, "An Improved K-means Clustering Algorithm, Communication Software and Networks (ICCSN), 2011 IEEE 3rd

International Conference on 27 (pp. 44-46), May 2011.

[12]. Canlas, Ruben. D. "Data mining in healthcare: Current applications and issues." School of Information Systems & Management, Carnegie Mellon University, Australia (2009).

[13]. 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, 2009, London, U.K.

[14]. Güvenir, H., Demiröz, G., &Ilter, N. "Learning differential diagnosis of erythemato-squamous diseases using voting feature intervals". Artificial Intelligence in Medicine, 13(3) 147-165, 1998.

[15]. Marty, Babu, G.P. and MN 1994. Clustering with evolution strategies Pattern Recognition, 27, 2, 321-329.

[16]. G K, G. Kesavaraj, Sukumaran, Surya, "A study on classification techniques in data mining", 2013 4th International Conference on Computing, Communications and Networking Technologies, ICCCNT-2013.

[17]. P. Tamilselvi and K. A. Kumar, "Unsupervised machine learning for clustering the infected leaves based on the leaf colours," 2017 Third International Conference on Science Technology Engineering & Management (ICONSTEM), pp. 106-110, Chennai, 2017.

[18]. G. Q. Wu, X. Wu, X. Zhu, and W. Ding, "Data mining with Big Data," Knowledge and Data Engineering, IEEE Transactions on, vol. 26, no 1, p. 97-107, 2014.