International Journal of Advanced Computer Technology

ISSN: 2319-7900 www.ijact.org Volume-X, Issue-V, October 2021

A Survey on Performance Improvement of Data Analysis Using Unsupervised K-Means Clustering

Arpana Kumari [1], Monika Raghuvanshi [2]
Computer Science and Engineering Department
Bhabha Engineering Research Institute, MP, Bhopal, India
[1]arpana7info@gmail.com, [2]monipriya21@gmail.com

Abstract: *The algorithms clustering implemented on the machines and made intelligent machines are called unsupervised machine learning algorithms. They can perform essential tasks by k-means clustering algorithm based on improved quantum particle swarm optimization algorithm is often more error in data analysis. As more data becomes available, more complex problems can be tackled and solved. The analysis of patient's data is becoming more critical to evaluate the patient's medical condition and prevent and take precautions for the future. With the help of technology and computerized automation of machines, data can be analyzed more efficiently. Managing the massive volume of data has many problems interrelated to data security. Experiments on actual datasets show that our technique will get similar results with standard ways with fewer computation tasks. Process mining and data mining techniques have opened new access for the diagnosis of disease.*

Similarly, data mining can provide effective treatment for a disease's triennial prevention; finally, an effective clustering result is obtained. The algorithm is tested with the UCI data set. The results show that the improved algorithm ensures the global convergence of the algorithm and brings more accurate clustering results.

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

I. INTRODUCTION

Data mining is the exploration and analysis of 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 the component of a more comprehensive process called knowledge discovery from a database [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 the 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 be used to cluster the input data in classes based on their statistical properties only. 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, the target variable has typically only been recorded for too small a number of 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 vast 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 various 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 describing the objects and their relationships. Their feature values can be used in many applications, such as knowledge discovery, vector quantization, pattern recognition, data mining, data dredging [5].

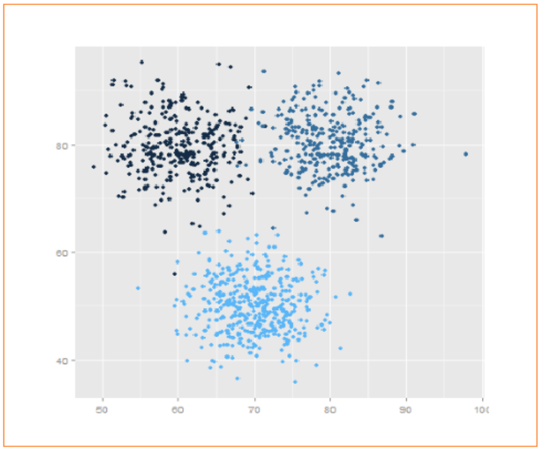

Figure 1 number of three clusters in clustering

There are mainly two techniques for clustering: hierarchical clustering and partitioned clustering. Data are not partitioned into a particular cluster in a single step. Still, 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. Each data object is in exactly one cluster [6].

II.RELATED WORK

L Bai et al. [7] find a better data clustering centre using a clustering algorithm to make the algorithm convergence faster and clustering results more accurate. A k-means clustering algorithm based on an 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 data set. The results show that the improved algorithm ensures the global convergence of the algorithm and brings 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, this paper proposes 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 before likewise as an unknown range of 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 selection of the initial centroid is better than choosing the initial centroids randomly.

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 evident 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 the fields of 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 represents an idea description by a group of feature intervals. A brand-new instance classification is predicated on a vote among the variety 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 gap is maintained. Thus, an interval could represent many categories by sorting the vote for every class. This formula has obtained 96% 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 reviewing 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.

III.EXPECT OUTCOME

Data mining uses a k-means clustering algorithm to create a cluster centre. Find minimum error of medical dataset analysis and best possible solution. The number of challenges of K-Mean clustering method based on unsupervised clustering algorithm using improve performance of dataset analysis, error minimizations using medical health care dataset analysis and best solution.

IV. CONCLUSION

Clustering algorithms are essential for extensive data analysis using unsupervised learning methods and may be thought of as an area of an overall data processing framework. Several algorithms were specifically designed to handle these problems, and k-means is concentrated on these problems, which may be self-addressed in the subsequent analysis. Medical data processing will facilitate arranging some strategies for identification and deciding activities. Data mining using k-means clustering-based cluster centre finds a minimum error of medical dataset analysis but gets a suboptimal solution. Clustering is the technique by which large datasets are divided into small data collections that are called clusters. A number of algorithms work well for clustering the data that can divide a dataset into clusters, it uses different breast cancer datasets from machine learning, and the algorithm is tested with UCI data set. Survey on K-Means clustering algorithm proposes different advantages and disadvantages in different K-Means application algorithm. In the survey, medical dataset processing will facilitate arranging some strategies for identification and deciding the level of disease activities and finding out centred on the utilization of unsupervised K-Means clustering based on machine learning for classification. Proposed clustering method based on the minimum error of medical dataset analysis.

REFERENCES

[1]. Xiaofan, Chen, and Wang Fengbin. "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 performance evaluation of human

resource." 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, Xiaolong Su, "An Improved K-means Clustering Algorithm, Communication Software and Networks (ICCSN), 2011 IEEE 3rd International Conference on 27 May 2011 (pp. 44-46).

[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 erythematous-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.