Clustering Performance Comparison using K-means Clustering Algorithm and IPCA

Yogiraj Singh Kushwah[1], Prof. Ashish Mohan Yadav[2]
Computer Science & Engineering Department
S. I. S. Tech, Bhopal, India
[1]yogiraj.singh.k@gmail.com, [2]ashishmohanyadav@sistec.ac.in

***ABSTRACT**: -* In the field of data mining is data analysis and supported unsupervised clustering algorithm. Unsupervised learning clustering one of the fastest growing research areas because of availability of the huge quantity of data analysis and extract useful information. Improve performance of unsupervised learning clustering and clustering an unsupervised classification that's the partitioning of a data set in a set of meaningful subsets and number of clusters. Unsupervised machine learning algorithm is based on extract information and mines the invisible information; meaningful data from the mountain of data, hidden patterns the finding out clusters. K-means clustering is one of the unsupervised machine learning strategies between all partitioning primarily based clustering strategies. The proposed algorithm is improving the performance of clustering algorithm (IPCA) bases on an experiment on the various dataset and find optimal solution. A proposed algorithm is minimizing error and optimization in the cluster and also the effectiveness of the proposed clustering algorithm. The dataset is then filtered for the empty values. The empty value set are then removed. The absolute expression is then identified form the dataset .This algorithm result in dataset clusters analysis as compare previous approach execution time is more as compare proposed approach and optimal number of cluster.

Keywords:- Clustering, K-Means Clustering ,Cluster Center, Partitioning Clustering, Machine learning ,Supervised Learning, Unsupervised Learning, Knowledge Discovery in Database.

I. INTRODUCTION

Data Mining is a process useful Knowledge finds in the large dataset. Different Data Mining technique used to mine the invisible, meaningful information from the mountain of the dataset. Data mining is term mine information and also another term as Knowledge Discovery in Database (KDD). The patterns based on they look for the Data Mining models and tasks are divided into two main categories Predictive models and Descriptive Models. While the Predictive Model is used to predict or possible to the feasibility of outcome, the other Descriptive model is used to describe the important features of the dataset. The predictive model is classification, regression, prediction or possible and time series analysis. Different models built-in in the descriptive model are clustering, summarization, Association rules and sequence discovery. Clustering is an also called an unsupervised learning. Clustering or cluster analysis can be defined as a data reduction tool used to generate different subgroups that are more manageable than the individual. Clustering is an also defined as a process used for organizing dataset or grouping of large dataset a large amount of information into meaningful groups or clusters based on similarity in data. Clusters are content similarly objects sets that have data similar on basis of common features and unlike to data in other clusters. Different applications and different areas where clustering plays an important role are machine learning dataset like image processing, data mining, marketing, text mining. Clustering and classifications are always confused but each other, both are separate conditions or term. Clustering is an unsupervised learning process for the reason that the resultant clusters are not known information before the execution which implies the absence of predefined classes in clustering. It is also called a classification is a supervised learning and in this process due to the presence of predefined classes. In this process, the high-quality clustering is to find high intra cluster similarity and low inter-cluster similarity [1].

II. Types of Machine Learning Algorithms

Machine learning is described as the acquisition of knowledge and the ability to use it. They explain that learning in data mining involves finding and describing structural patterns in data for the purpose of helping to explain that data and make predictions from it. For example, the data could contain examples of customers who have switched to another service provider in the telecommunication industry and some that have not. The output of learning could be the prediction of whether a particular customer will switch to another service provider or not. There are two common machine learning types of learning, first supervised learning and second unsupervised learning. Supervised learning in this training data it contributes each the input and also the desired results. These strategies are quick and correct. The right results are known and are given in inputs to the model during the learning method. Supervised models are neural network, several layers Perception, decision trees. Unsupervised learning in the model isn't maintained with the correct results throughout the training. It may be wont to cluster the input file in categories on the support of their probability properties only. Unsupervised models are non identical types of clustering, amplitude and normalization, k-means, self organizing maps [2]

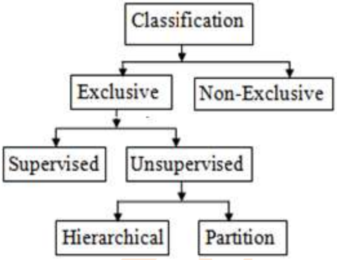

Figure 1 Types of Classification

Supervised: - Supervised are classified on the basis of supervised learning. In supervised learning external knowledge or information is provided. Here each example is a pair consisting of an input object and a desired output value. A supervised algorithm analyses the training data and produces an inferred function which can be used for mapping other examples. The information or data (observations, measurements, etc.) are labeled with pre-defined classes. It is like that a "teacher" gives the classes (supervision).Like human learning from past experiences, a computer does not have "experiences". A computer system learns from data, which represent some "past experiences" of an application domain. Our focus is to learn a target function that can be used to predict the values of a discrete class attribute, e.g., approve or not-approved, and high-risk or low risk. The assignment is normally called supervised learning, classification, or inductive learning [3].

Unsupervised: - Entirely without reference to external information. It can be achieved through clustering. A way of grouping together data samples that is similar in some way according to some criteria they just pick out. So, it's a method of data exploration, a way of looking for patterns or structure in the data that are of interest. It involves the use of descriptors and descriptors extraction. Descriptors are set of words that describe the contents within the cluster. It is considered to be a centralized process. E.g.: Web document clustering for search users. It requires no predefined classes or category. In unsupervised clustering, they have unlabelled gathering of documents. The objectives are to cluster the papers without extra information or interference such that papers within a cluster are more similar than papers between clusters. Usual Clustering method can be types into two main class as partitioned and Hierarchical. Here they discuss these groups and their main representatives. Unsupervised clustering, they have unlabelled collection of documents. The plan is to cluster the documents without extra information or intervention such that documents within a cluster are more similar than documents between clusters. Traditional clustering techniques can be categorized into two major groups as partition and hierarchical. In this chapter they discuss these groups and their main representatives [4]. In unsupervised clustering, they have unlabelled gathering of documents. The objectives are to cluster the papers without extra information or interference such that papers within a cluster are more similar than papers between clusters. Usual Clustering method can be types into two main class as partitioned and Hierarchical. Here they will be explaining these class and their main representatives.

Partition Clustering Techniques: - This algorithm manufacture un-nested, non-overlapping partitions of documents that typically domestically optimize a cluster criterion. The overall methodology is as follows: given the quantity of clusters k, an initial partition is constructed; next the cluster answer is refined iteratively by moving documents from one cluster to a different. within the following sub-sections are discuss the foremost common partition formula k-means, and its variant bisecting k-means that has been applied to cluster documents by Steinbach et al. in and has been shown to typically trounce agglomerate hierarchical algorithms [5].

K-Means Clustering: - the thought behind the k-means formula is that every of k clusters may be described by the mean of the documents allotted thereto cluster, that is named the centroid of that cluster. There are 2 versions of k-means formula famous. The primary version is that the batch version and is additionally referred to as Forgy's formula. It consists of the subsequent two-step major iterations: (i) assign all the documents to their nearest centroids (ii) Recomputed centroids of freshly assembled teams before the iterations begin, first k documents are chosen because the initial centroids. Iterations continue till a stopping criterion like no reassignments occur is achieved. Initially, k documents from the corpus are chosen at random because the initial centroids. Then, iteratively documents are allotted to their nearest centroid and centroids are updated incrementally, i.e., once every assignment of a document to its nearest centroid. Iterations stop, once no reassignments of documents occur.

Bisecting K-Means: - though bisecting k-means is really a divisive cluster formula that achieves hierarchy of clusters by repeatedly applying the fundamental k-means formula, they discuss it during this section because it could be a variant of k-means. In every step of bisecting k-means a cluster is chosen to be split and it's split into 2 by applying basic k-means for k = two. the biggest cluster, that's the cluster containing the most variety of documents, or the cluster with the smallest amount overall similarity may be chosen to be split [6].

III. LITERATURE REVIEW

Wei Du et al. [7]. "A New Projection-based K-Means Initialization Algorithm".K-means is widely used in many areas for the features of its efficiency and easily understood. However, it is well known that the K-Means algorithm may get suboptimal solutions, depending on the choice of the initial cluster centers. As a partition based clustering algorithm, K-Means is widely used in many areas for the features of its efficiency and easily understood. However, it is well known that the K-Means algorithm may get suboptimal solutions, depending on the choice of the initial cluster centers.

Sijia Liu et al. [8] described a useful survey of fuzzy clustering in main three categories. Primary class is basically the fuzzy clustering depends on exact fuzzy relation. The next one is the fuzzy clustering based on single purpose function. Finally, it is given an overview of a nonparametric classifier. That is the fuzzy common used a k nearest neighbor rule. The fuzzy clustering algorithms have obtained great achievement in a variety of substantive areas.

Faming LU al. [9] plan a simple and qualitative method and k means clustering method to classify and using NBA guard and Euclidean distance as a measure of similarity distance. This work display by using k-Means clustering algorithm and 120 NBA guards' data. Manual classification of traditional methods is improved using this model. According to the existing statistical data, the NBA players are classified to make the classification and evaluation objectively and scientifically. This work shows that this is very effective and reasonable methodology. Therefore, based on classification result the guards' type can be defined properly. Meanwhile, the guards' function in the team can be evaluated in a fair and objective manner.

Liu Xumin et al. [10] present the analysis of shortcomings of the standard k-means algorithm. As k-means algorithm has to calculate the distance between each data object and all cluster centers in iteration. This repetitive process affects the efficiency of clustering algorithm. An improved k-means algorithm is proposed in this paper. A simple data structure is required to store some information in iteration which is to be used in the next iteration. Computation of distance in iteration is avoided by the proposed method and saves the running time. The work of research paper shows that planned method can effectively get better the speed and accuracy of clustering, reducing the computational complexity of the k- means.

Madhu Yedla et al. [11]. Improving K-means Clustering Algorithm and cluster data analysis is one of the primary data analysis methods and k-means is one of the most well known accepted clustering algorithms. The k-means algorithm is one of the frequently used clustering methods in data mining, due to its performance in clustering massive data sets. The last cluster end answer of the k-means clustering algorithm very much depends upon the correctness of the initial centroids, which are selected randomly. The original k-means algorithm converges to local minimum, not the global optimum. Many improvements were already proposed to improve the performance of the k-means, but most of these require additional inputs like threshold values for the number of data points in a set. In this search paper a new method is proposed for discovery the better initial centroids and to make available an efficient way of assigning the data points to suitable clusters with reduced time complexity.

Kevin Dong et al. [12], "Assessing Pancreatic Cancer Risk Associated with Dipeptidyl Peptidase 4 Inhibitors: Data Mining of FDA Adverse Event Reporting System (FAERS)" Identifying cancer risks related to medicative agents plays a crucial role in cancer management and hindrance. Case reports of cancers related to pharmacotherapy are escalating within the Food and Drug Administration Adverse Event coverage System (FAERS).

M S Mahmud et al. [13] gave an algorithm to compute better initial centroids based on heuristic method. The recently existing algorithm outcome in very much accurate clusters with decrease in computational time. In this method used firstly compute the standard score of each data points that consists of multiple attributes and weight factor. Sort is applied to sort the output that was previously generated. Different data points are then divided into different k cluster i.e. number of desired cluster. To conclude the nearest possible information or data point of the mean is taken as original centroid new outputs show that the algorithm reduces the number of iterations to assign data into a cluster. The problem of assigning number of desired cluster as input.

Jaspreet K S et al. [14]. Enhanced the traditional k-means by introducing Ranking method. Author introduces Ranking Method to overcome the deficiency of more execution time taken by traditional k-means. The position technique is a way to discover the rate of similar data and to get better search effectiveness. The tool used to implement the improved algorithm is Visual Studio 2008 using C#. The benefits of k-means are also analyzed the author finds k-means as fast, robust and easy understandable algorithm.

Xiaolong Su et al. [15] discuss an improved k-means clustering algorithm to deal with the problem of outlier detection of existing k-means algorithm. The planned method uses noise data or information filter to deal with this problem. Compactness based outlier discovery technique is applied on the data to be clustered so as to eliminate the outliers.

IV. SIMULATION AND RESULT ANALYSIS

Matlab Toolbox: The Statistics Toolbox is a collection of tools built on numeric computations. The the MATLAB for performing toolbox supports a wide range of statistical tasks, ranging from random number generation, to curve fitting, to design of experiments and statistical process control. The matlab statistical toolbox are provides a tool types building-block probability and mathematical functions and graphical, interactive tools. The first type of function support can be called from the rule or from your own applications. They can show the MATLAB code and use these functions, change the method any toolbox function works by repetition and renaming The M-file, then modifying your copy and even extend the toolbox by adding our own M-files. Secondly, the toolbox provides a number of interactive tools that enables us to access functions through a graphical user interface In order to test the proposed method, Simulation using MATLAB are performed on input images. MATLAB (matrix laboratory) is a multi paradigm numerical computing situation and 4th generation programming language. It is developed by math tool in MATLAB. MATLAB work allows matrix strategy, plotting of function and data, implementation of algorithm, construction of user interfaces and interfacing with programs. Identify various challenges in the field of data mining and following objective in unsupervised clustering algorithm. Useful information extracts in clusters Increase accuracy in clustering technique. Extract reliable data. Minimize error-values in clustering.

1. Bacteria_etec dataset based result graph show below. Dataset clusters analysis as compare previous approach but execution time is more as compare previous approach. K-mean algorithm is time take minim as compare to IPCA.

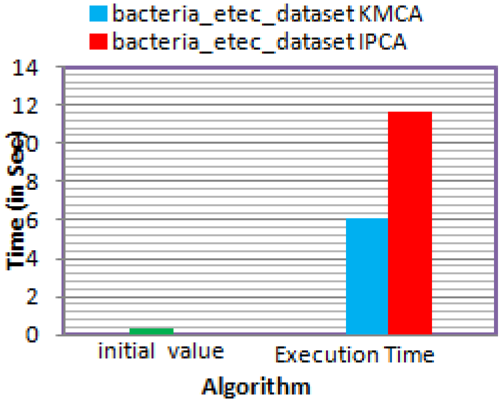

Figure2 In case of Bacteria_etec dataset analysis execution processing time comparison between K-mean & IPCA.

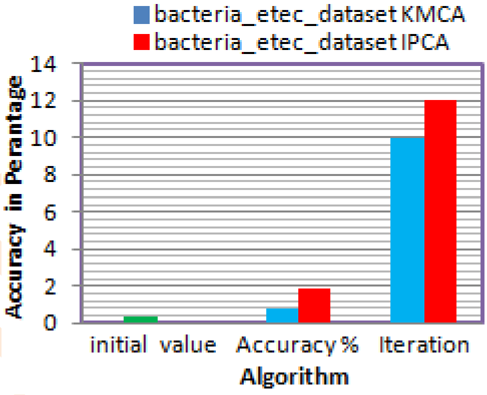

Figure3 In case of Bacteria_etec data analysis accuracy comparison between K-mean & IPCA.

2. Accuracy: Bacteria_etec dataset based result graph show below. K-mean algorithm is more error in dataset clustering as compare to IPCA .IPCA is better as compare to k-means because k-mean dataset cluster error rate is more but IPCA is minim dataset cluster error. At the time of IPCA proceed of error free data is greater than the previous approach k-mean, graph shows minimum error dataset clusters as compare previous approach.

V. CONCLUSION

Data clustering approach IPCA has been proposed. Comparative analysis shows that proposed approach have better convergence to lower quantization errors, same execution time. Accuracy measurement signifies the real impact of the proposed algorithm; proposed approach is more accurate. IPCA more accurate as compare K-Means Algorithm. If done more efficiently then execution time can be reduced is proposed work. The result analysis shows that the performance of the IPCA improves with the application of K-mean clustering algorithm as compared to different data set. The improvement within the error clearly dataset that the IPCA performs increased as compared to K-mean clustering on database or data set. Result analysis processes used two database graves thairaid and WPBC Dataset. Wisconsin prognostic breast cancer dataset based result graph show below .K-mean algorithm is more error in dataset clustering as compare to IPCA .IPCA is better as compare to k-means because k-mean dataset cluster error rate is more but IPCA is minim dataset cluster error. K-mean algorithm is time take mini as compare to IPCA. The graph shows above with the purpose algorithm high accuracy as compared to the normal previous algorithm K-Mean has because the IPCA data set has less error in dataset also called filtered data set.

REFERENCES

[1]. Wang Shunye "An Improved K-means Clustering Algorithm Based on Dissimilarity" 2013 International Conference on Mechatronic Sciences, Electric Engineering and Computer ,Dec 20-22, Shenyang, China IEEE, 2013.

[2]. Burges, C. J. C., "A Tutorial on Support Vector Machines for Pattern Recognition", Data Mining and Knowledge Discovery, Vol. 2, No. 2, pp. 121–167, 1998.

[3]. N. Suguna, and Dr. K. Thanushkodi, "An Improved k-Nearest Neighbor Classification Using Genetic Algorithm", IJCSI International Journal of Computer Science Issues, Vol. 7, Issue 4, No 2, July 2010.

[4]. Shailesh S. Deshpande and Girish Keshav Palshikar and G. Athiappan "An unsupervised approach International Conference on Management of Data COMAD 2010.

[5]. Steinbach, M., G. Karypis, and V. Kumar, "A Comparison of Document Clustering Techniques", KDD Workshop on Text Mining, 1999.

[6]. D Ramesh, B Vishnu Vardhan, "Data Mining Techniques and Applications to Agricultural Yield Data", International Journal of Advanced Research in Computer and Communication Engineering Vol. 2, Issue 9, September 2013.

[7]. Wei Du, Hu Lin, Jianwei Sun, Bo Yu and Haibo Yang, "A New Projection-based K-Means Initialization Algorithm", IEEE ,August 12-14, China, 2016.

[8]. Don Kulasiri, Sijia Liu, Philip K. Maini and RadekErban, "Diffuzzy: A fuzzy clustering algorithm for complex data sets", International Journal of Computational Intelligence in Bioinformatics and Systems Biology vol.1, no.4,pp. 402-417, 2010.

[9]. Libao ZHANG, Faming LU, An LIU, Pingping GUO, Cong LIU, "Application of K-Means Clustering Algorithm for Classification of NBA Guards", International Journal of Science and Engineering Applications Volume 5 Issue 1, ISSN- 2319-7560, 2016.

[10]. Shi Na, Liu Xumin, Guan Yong, "Research on K-means Clustering Algorithm: An Improved K-means Clustering Algorithm", Intelligent Information Technology and Security Informatics, 2010 IEEE Third International Symposium on, (pp. 63-67),2-4 April, 2010.

[11]. Madhu Yedlz, Srinivasa Rao Pathakota, T M Srinivasa, "Enhancing K-means Clustering Algorithm with Improved Initial Center", IJCSIT, Vol. 1 (2), 2010.

[12]. Xiaodong Feng, Amie Cai, Kevin Dong, Wendy Chaing, Max Feng, Nilesh S Bhutada, John Inciardi", Assessing Pancreatic Cancer Risk Associated with Dipeptidyl Peptidase 4 Inhibitors: Data Mining of FDA Adverse Event Reporting System (FAERS)",J Pharmacovigilance ,2013.

[13]. Md. Sohrab Mahmud, Md. Mostafizer Rahman, and Md.Nasim Akhtar "Improvement of K-means Clustering algorithm with better initial centroids based on weighted average",7th International Conference on Electrical and Computer Engineering 20-22 December, Dhaka, Bangladesh, IEEE, 2012.

[14]. Jaspreet Kaur Sahiwal, Navjot Kaur, Navneet Kaur "Efficient K-means clustering Algorithm Using Ranking Method In Data Mining", ISSN: 2278 – 1323 International Journal of Advanced Research in Computer Engineering & Technology Volume 1, Issue 3, May2012.

[15]. Juntao Wang & Xiaolong Su, "An improved K-Means clustering algorithm", IEEE, 2011.