A Survey on Unsupervised Clustering Algorithm based on K-Means Clustering

Yogiraj Singh Kushawah
M. Tech. Scholar Depart. of CSE
SISTech, R.G.P.V, Bhopal
M.P., India

Ashish Mohan Yadav
Department of CSE
SISTech, R.G.P.V, Bhopal
M.P., India

ABSTRACT
Data mining are data analysis supported unsupervised clustering algorithm is one of the quickest growing research areas because of availability of huge quantity of data analysis and extract usefully information based on new improve performance of clustering algorithm. Clustering is an unsupervised classification that's the partitioning of a data set in a set of meaningful subsets .Machine learning is based on extract and mine the invisible, meaningful data from mountain of data, hidden patterns the finding out clusters may be a supported unsupervised learning. K means is one of the best unsupervised learning strategies among all partitioning primarily based clustering strategies. The proposed algorithm is improving performance of clustering algorithm (IPCA) bases on experiment on various dataset. A proposed algorithm is minimizing error and optimization in cluster and also the effectiveness of the proposed clustering algorithm.

Keywords
Clustering, K-means clustering cluster center, partitioning clustering, unsupervised learning.

1. INTRODUCTION
Data mining is that the process of extracting useful and hidden information or data from large datasets. the data therefore extracted will be used to improve an unsupervised clustering .Clustering is an unsupervised technique that groups the knowledge of similar objects with minimum cluster distance into a cluster by eliminating the inappropriate data objects .Data mining consists of six basic kinds of tasks that are Anomaly detection, Association rule learning, Clustering, Classification, Regression and report. Clustering is one of the necessary tasks of data mining. Clustering is that the unsupervised classification of data objects into groups or clusters. [1]. kinds of clusters are well- separated cluster, contiguous cluster, shared property or conceptual cluster, center-based cluster and density primarily based cluster. clustering discovers the close groups within the information and helps to find the data discovery in engineering and scientific domains like psychology, medicine, remote sensing, etc. hierarchical clustering, Partition based mostly group and grid based clustering are the categories of clustering algorithm. The partitional cluster algorithms, that disagree from the hierarchical clustering algorithms, are typically to make some sets of clusters at begin and partition the information into similar groups when every iteration. Partitional clustering is additional used than hierarchical clustering as a result of the dataset are often divided into quite two subgroups during a single step except for hierarchy technique, invariably merge or divide into two subgroups. Clustering is of two types particularly, laborious clustering and soft clustering. Every component during a sample belongs to exactly one cluster comes below hard bunch. Every component belongs to every cluster during a population refers to soft bunch. Data processing is that the development of evaluating data from totally different views and summarizing it into helpful info. Data processing consists of extract, transfer, and load group action data onto the information warehouse system, Store and manages the information during a three-dimensional info system. Data processing involves the anomaly detection, association rule learning, classification, regression, summarization including clustering. In this paper, clustering analysis is completed. Cluster Analysis, an automatic procedure to find similar objects from a database. It's an elementary operation in data processing [2].In data mining the data is mined using two learning.

- Supervised learning
- unsupervised clustering

1.1.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.

1.2.Unsupervised Learning:
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. [3]

2. CLUSTERING
Data mining using in clustering could be a set of such clusters, sometimes containing all objects within the data set. To boot, it should determine the connection of the clusters to every different, as an example, a hierarchy of clusters embedded in one another. The notion of a "cluster" cannot be altogether defined, that is one in all the explanations why there are such a large amount of clustering algorithms. There's a familiar denominator: a group of data objects. However, completely different researchers use different cluster models, and for every of those cluster models once more dissimilar algorithms may be given. The notion of a cluster, as found by dissimilar algorithms, varies significantly in its properties. Understanding these "cluster models" is essential to understanding the variations between the assorted algorithms Cluster analysis is an iterated method of data discovery and it's a variable statistical technique that identifies groupings of the info objects supported the inter-object similarities computed by a selected distance metric .Clustering algorithms may be classified into two categories: hierarchical clustering and Partitional clustering[4,5].

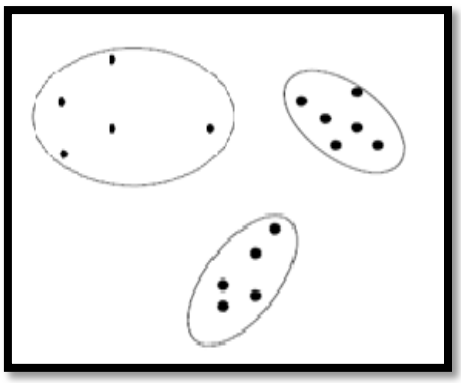

Fig1: Clustering

2.1. Types of Clusters [6]

2.1.1. Well-separated clusters:

A cluster could be a settle of points so any point in a very cluster is nearest (or a lot of similar) to each different point within the cluster as differentiate to the other point that's not within the cluster.

2.1.2. Center-based clusters

A cluster could be a settle of objects specified an object in a very cluster is highest (more similar) to the "center" of a cluster, than to the middle of the other cluster. The center of a cluster is commonly a centroid.

2.1.3. Contiguous clusters

A cluster could be a position of points so a point in a very cluster is nearest (or a lot of similar) to at least one or additional different points within the cluster as differentiate to any point that's not within the cluster.

2.1.4. Density-based clusters

A cluster could be a heavy region of points that is split by according to the low-density regions, from different regions that's of high density.

3. LITERATURE REVIEW

Tajunisha et al. [7]. Performance Analysis of k-Means with different initialization methods for high dimensional data .uses Principal Component Analysis (PCA) for dimension reduction and to find initial cluster centers. The variable with the highest Eigen value calculated using PCA is taken as first principal component along which partitioning is done, on the basis of which k subsets are formed and k median values are taken as initial k centers.

Bouhmala et al. [8]. Combined Genetic Algorithm and K-Means to improve the quality of clusters formed and speed up their search process. The performance of GAKM is tested over the datasets such as iris, glass, etc., and that has been taken from Machine learning repository. The experimental results have proved that GAKM converges faster while comparing to standard Genetic Algorithm. Though this algorithm failed to capture the best quality of clusters, it is unsuitable for the maximizing both homogeneity and heterogeneity within same clusters and with different clusters respectively.

Rouhollah et al. [9]. Projected a model called GA clustering for improving K-means algorithm. The algorithm has been performed on well known datasets namely iris, crude oil. The experimental results provide clustering standard μ, the lower the value of μ gives the better cluster formation of data compared to traditional K-means clustering algorithm.

N. Kaur et al. [10]. Enhanced the traditional kmeans by introducing Ranking method. Author introduces Ranking Method to overcome the deficiency of more execution time taken by traditional k-means. The Ranking Method is a way to find the occurrence of similar data and to improve search effectiveness. The tool used to implement the improved algorithm is Visual Studio 2008 using C#.The advantages of k-means are also analyzed in this paper. The author finds k-means as fast, robust and easy understandable algorithm. He also discuss that the clusters are non-hierarchical in nature and are not overlapping in nature. The process used in the algorithm takes student marks as data set and then initial centroid is selected. Euclidean distance is then calculated from centroid for each data object. Then the threshold value is set for each data set. Ranking Method is applied next and finally the clusters are created based on minimum distance between the data point and the centroid. The future scope of this paper is use of Query Redirection can be used to cluster huge amount of data from various databases.

Sandeep Rana et al. [11]. Proposed a new Hybrid Sequential clustering approach. They have used PSO and K-Means algorithm in sequence for data clustering. This approach was proposed to overcome the drawbacks of both algorithms as well as improves clustering and avoids being stagnated. Four kinds of data sets have been tested in order to obtain comparative results. For comparison purpose, different algorithms such as PSO, K-Means, Hybrid K-Means PSO, and Hybrid K-Means + Genetic Algorithm were considered. The proposed algorithm generates more accurate and robust clustering results.

Bara'a Ali Attea et al. [12]. Discovered that performance of clustering algorithms degrades with more and more overlaps among clusters in a data set. These facts have motivated to develop a fuzzy multi-objective particle swarm optimization framework (FMOPSO) in an innovative fashion for data clustering, which is able to deliver more effective results than state-of-the-art clustering algorithms. To ascertain the superiority of the proposed algorithm, number of statistical tests has been carried out on a variety of numerical and categorical real life data sets.

Chetna Sethi et al. [13]. Proposed a Linear PCA based hybrid K-Means clustering and PSO algorithm (PCA-K-PSO). In (PCA-K-PSO) algorithm the fast convergence of K-Means algorithm and the global searching ability of Particle Swarm Optimization (PSO) are combined for clustering large data sets using Linear PCA. Better clustering results can be obtained with PCA-K-PSO as compared to ordinary PSO. This was effectively developed in order to make its use for efficient clustering of high- dimensional data sets.

Rui Xu et al. [14]. "Survey of Clustering Algorithms" .Data analysis plays an indispensable role for understanding various phenomena. Cluster analysis, primitive exploration with little or no prior knowledge, consists of research developed across a wide variety of communities. The diversity, on one hand, equips us with many tools. On the other hand, the profusion of options causes confusion. a survey clustering algorithms for data sets appearing in statistics, computer science, and machine learning, and illustrate their applications in some benchmark data sets, the traveling salesman problem, and bioinformatics, a new field attracting intensive efforts. Several tightly related topics, proximity measure, and cluster validation, are also discussed.

Xiaohui et al. [15]. Used Particle Swarm Optimization (PSO) for document clustering. K-Means algorithm is most commonly used partitioning algorithm for clustering large datasets but it produces local optimal solution. In contrast to localized searching property of K-Means, PSO performs globalized search using entire solution space. Authors used PSO, K-Means and hybrid PSO clustering algorithm on four document datasets which are derived from Text Retrieval Conference (TREC) and contains 414, 313, 204, 878 documents respectively. In hybrid PSO two modules are used the PSO module and the K-Means module. In each experiment PSO and K-Means run 100 iterations while in hybrid PSO approach, PSO algorithm is executed for 90 iterations and then K-Means is executed for 10 iterations.

Tapas Kanungo et al. [16]. An efficient k-means clustering algorithm: analysis and implementation .In k-means clustering, they are given a set of n data points in d-dimensional space Rd and an integer k and the problem is to determine a set of k points in Rd, called centers, so as to minimize the mean squared distance from each data point to its nearest center. A popular heuristic for k-means clustering is Lloyd's algorithm. In this paper, present a simple and efficient implementation of Lloyd's k-means clustering algorithm, which they call the filtering algorithm. This algorithm is easy to implement, requiring a kd-tree as the only major data structure. A establish the practical efficiency of the filtering algorithm in two ways. First, present a data-sensitive analysis of the algorithm's running time, which shows that the algorithm runs faster as the separation between clusters increases. Second, present a number of empirical studies both on synthetically generated data and on real data sets from applications in color quantization, data compression, and image segmentation.

4. EXPECT OUTCOME

Identify various challenges in the field of data mining and following objective in unsupervised clustering algorithm.

1. useful information extract in clusters
2. Increase accuracy in clustering technique.
3. Extract reliable data
4. Minimize error-values in clustering and best possible answer

5. CONCLUSION

Data mining is to divide clusters from large data set and transform it into an understandable extract information .Clustering plays a very important task in data mining applications and data mining analysis. Clustering can be done by different algorithms such as grid based algorithm, hierarchical algorithm, partitioning algorithm and density based algorithm. Grid based clustering has a finite number of cells that form grid structure. Hierarchical based clustering is connectivity based clustering. Partitioning based algorithm is the centroids based clustering. These clustering techniques produce efficient clusters when compare to other clustering with low cost.

6. REFERENCES

[1] R.S. Santos, S.M.F. Malheiros, S. Cavalheiro, J.M. Parente de Oliveira, "A Data Mining system for providing analytical information on Brain tumors to public health decision makers", Computer Methods And Programs in Biomedicine, ISSN:0169-2607, pp. 296-282, 2013.

[2] Jirong Gu, Jieming Zhou,Xianwei Chen, "An Enhancement of K-means Clustering Algorithm", in proceeding of Business Intelligence and Financial Engineering ,International Conference, 2009, Key Lab. of the Southwestern Land Resources Monitoring, Sichuan Normal Univ., Chengdu, China, 2009.

[3] Amandeep Kaur Mann, Navneet Kaur,"Survey Paper on Clustering Techniques"

[4] J. Han and M. Kamber, Data Mining: Concepts and Techniques, Morgan Kaufmann Publishers, San Francisco, CA, 2001

[5]https://en.wikipedia.org/wiki/Cluster_analysis.

[6] Kapil Joshi, Himanshu Gupta, Prashant Chaudhary, Punit Sharma, "Survey on Different Enhanced K-Means Clustering Algorithm", International Journal of Engineering Trends and Technology (IJETT) – Volume 27 Number 4 - September 2015.

[7] Tajunisha and Saravanan, "Performance Analysis of k-Means with different initialization methods for high dimensional data" International Journal of Artificial Intelligence & Applications (IJAIA), Vol.1, No.4, October 2010 .

[8] N. Bouhmala, A. Viken, J. B. Lonnum, "Enhanced Genetic Algorithm with K-Means for the Clustering Problem", International Journal of Modeling and Optimization, pp. 150-154, 2015.

[9]Rouhollah Maghsoudi, Arash Ghorbannia Delavar, Somayye Hoseyny, Rahamatollah Asgari, Yaghub Heidari, "Representing the New Model for Improving K-meansClustering Algorithm based on Genetic Algorithm", The Journal of Mathematica and Computer Science, pp. 329-336, 2011.

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

[11] Sandeep Rana, Sanjay Jasola, and Rajesh Kumar, "A hybrid sequential approach for data clustering using K-Means and particle swarm optimization algorithm," International Journal of Engineering, Science and Technology Vol. 2, No. 6, 2010, pp. 167-176.

[12] Bara'a Ali Attea, "A fuzzy multi-objective particle swarm optimization for effective data clustering," Springer-July 2010, pp. 305-312.

[13] Chetna Sethi and Garima Mishra, "A Linear PCA based hybrid K-Means PSO algorithm for clustering large dataset," International Journal of Scientific & Engineering Research, Volume 4, Issue 6, June-2013, pp.1559-1566.

[14] Rui Xu, and Donald Wunsch , "Survey of Clustering Algorithms" ,EEE Transactions On Neural Networks, Vol. 16, No. 3, May 2005.

[15] C. Xiaohui, E. P. Thomas and P. Paul, "Document Clustering using Particle Swarm Optimization", Swarm Intelligence Symposium, pp. 185-191, Pasadena, CA, USA, 2005.

[16] Tapas Kanungo, 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 And Machine Intelligence, Vol. 24, No. 7, July 2002.