ISSN: 2278 – 909X
International Journal of Advanced Research in Electronics and Communication Engineering (IJARECE)
Volume 7, Issue 6, June 2018

A REVIEW: OPTIMAL CLUSTERINGUSING K-MEANS GENETIC ALGORITHMS IN DATA MINING

Monika, Nisha Pandey

Shri Ram college of Engineering and Management

Abstract

The K-means method is one of the most widely used clustering methods and has been implementedin many fields of science and technology. One of the major problems of the k-means algorithm is that it may produce empty clusters depending on initial center vectors. Genetic Algorithms (GAs) are adaptive heuristic search algorithm based on the evolutionary principles of natural selection and genetics. In this paper we present surveyon k-means algorithm and GA that efficiently eliminates this empty cluster problem.

Keywords: Cluster Analysis, Genetic Algorithm, K-Means etc.

I. INTRODUCTION

Data mining is the analysis of data and the use of software techniques for finding patterns and regularities in the set of data [1]. Data mining involves the use of sophisticated data analysis tools to discover previously unknown, valid pattern and relationship in large data set [2]. Data mining, a synonym to "knowledge discovery in databases" is a process of Analyzing data from different perspectives and summarizing it into useful information. Clustering [3] is useful technique for the discovery of data distribution and patterns in the underlying data. Clustering is an example of unsupervised classification. Classification refers to a procedure that assigns data objects to a set of classes. Unsupervised Classifications means that clustering does not depend on predefined classes and no external teacher set is used. In clustering we measure the dissimilarity between objects by measuring the distance between each pair of objects. These measure include the Euclidean, Manhattan and Minkowski distance clustering method convert that information into various clusters where object in that cluster having similar proper-ties as compare to other but not same to other clusters properties. There are various efficient techniques used to solve the problem for large data clustering. Clustering techniques and implementation used for getting scalability and performance in such data analysis. By using cluster analysis techniques it is very easy to handle complex data sets and K-means is widely used for producing clusters in many application. It is also used for automatically organized data, compression form and finding some hidden structure [4][5][6]. In the figure 1 shows basic cluster formation from given dataset when we applying K-means algorithm. In first stage by selecting random initial centroid and clusters the data object in the dataset. Now in stage second recalculating centroid from first iteration due to this as figure shows some of the data object move from one cluster to another. In third stage of the figure centroid remain constant which means convergence is found. In this way all data object is clustered as respective cluster hence selection of initial centroid is main task in cluster formation

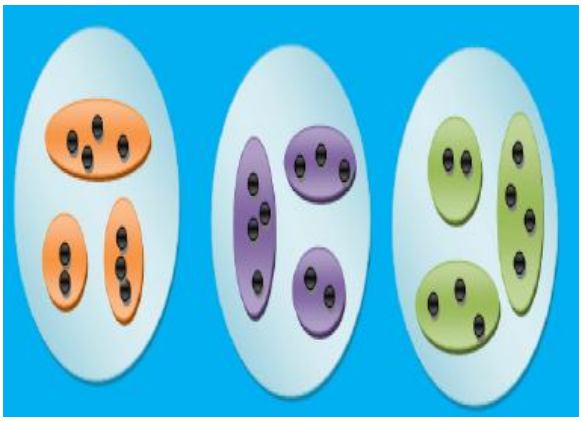


Figure 1: Depiction of Cluster Formation

A. Requirements of Clustering in Data Mining

Here are the typical requirements of clustering in data mining:

Scalability - We need highly scalable clustering algorithms to deal with large databases.

Ability to deal with different kind of attributes - Algorithms should be capable to be applied on any kind of data such as interval based (numerical) data, categorical, binary data.

Discovery of clusters with attribute shape - The clustering algorithm should be capable of detect cluster of arbitrary shape. That should not be bounded to only distance measures that tend to find spherical cluster of small size.

High dimensionality - The clustering algorithm should not only be able to handle low- dimensional data but also the high dimensional space.

Ability to deal with noisy data - Databases contains noisy, missing or erroneous data. Some algorithms are sensitive to such data and may lead to poor quality clusters.

Interpretability - The clustering results should be interpretable, comprehensible and usable.

II. HYBRID GA-K-MEANS CLUSTERING

K-means clustering is an unsupervised pattern classification method, applied directly to industrial environments without the need to being trained by data measured on a machine under a fault condition. Further advantage of this method is its ease of programming. The K-means algorithm seeks to partition the data into *K* groups or clusters so that the within-group sum of squares is minimized; that is, it seeks the cluster centers $\{\mu j, j=1,\ldots,k\}$ that minimize:

$$M = \sum_{j=1}^{k} S_j \quad (1)$$

Where the within-group sum of squares for group *j* is:

$$S_j = \sum_{i=1}^{n} z_{ji} \left| x_i - \mu_j \right|^2 \quad (2)$$

in which zj_i=1 if *xi* is in group *j* (of size $n_j = \sum_{i=1}^{n} z_{ji}$) and zero otherwise; $\mu j$is the mean of group *j*,

$$\mu_j = \frac{1}{nj} \sum_{i=1}^{n} z_{ji} x_i \quad (3)$$

The first step in the application of K-means clustering is to find a set of initial centers. K-means clustering is an iterative hill-climbing algorithm and it is significantly sensitive to the initial randomly selected cluster centers. Varying the starting conditions can produce different stable cluster. Although the K-means algorithm had been applied to many practical clustering problems successfully, it may converge to a partition that is significantly inferior to the global optimum [8].

Clustering by the means of GA seems to overcome the problem of hill climbing algorithms in clustering. Commonly used hill climbing algorithms can only generate a local optimal solution; whereas, meta-heuristic algorithms such as GA are able to escape from local optima with the help of mutation operator. In the following, GA clustering, used in this paper, is described in details:

A. Population Initialization

GA performs on the population of potential solutions. Each individual of this population is named as "chromosome". Each chromosome of the GA represents the centers of the clusters and it is a vector with $K \times N$ columns; where *K* is the number of clusters and *N* is the dimension of each clusters. The *K* cluster centers encoded in each chromosome are initialized to *K*randomly chosen points from the data set. This process is repeated for each of the chromosomes in the population.

B. Fitness function

The clustering metric M of each individual is calculated by (1), by which the sum of the distance of each data in the training set is calculated from its cluster center. In this article,
Euclidean distance is used as the distance metric.

C. Crossover

In this article, single point crossover with a fix crossoverprobability *pc* is adapted. For each pair of chromosomes thatundergoes the crossover operation, a random integer, called the crossover point, is generated in the range [*1*, *K-1*]. The portions of the chromosomes lying to the right of the crossover point are exchanged to produce two offspring.

D. Mutation

Each chromosome undergoes mutation with a fixed probability *pm*. In order to perform mutation, for each selected individual, a mutation point is determined randomly and its value is exchanged with another randomly generated value in the variable's range. Mutation operator performs the global search in the search space as an effort to escape from local optimum.

After finding the cluster centers from training data with either K-means or GA-K-means, the classification accuracy using K-Nearest Neighbor (KNN) method on the test data will be reported in addition to distance metric from (1). The accuracy measure is calculated as the following:

$$Accuracy = \frac{\sum_{i=1}^{n} \delta(y_i, c_i)}{n} \quad (4)$$

Where*n* is the number of samples in the testing data, *yi*and *ci* denote the true category label and the obtained cluster labelwith either methods, respectively. *δ(y,c)* is a function thatequals 1 if *y=c* and equals 0 otherwise [9].

Further improvements can be achieved by using the hybrid method of GA and K-means. In other words, in hybrid GA and K-means (GA-K-means), GA have applied for predefined iterations at first and then the best solution of GA is chosen to be the initial point for K-means clustering.

III. LITRATURE SURVEY

Zheng B, Yoon S. W. and Lam S. S. in 2014 purposed [7] a hybrid of K-means algorithm and support vector machine algorithm for feature extraction. K-means is used for recognizing the hidden patterns of tumor. The tumor feature data set is classified into malignant (cancerous) and benign (aren't cancerous) sets separately. The membership function is used to find the similarity between incoming tumor and symbolic tumor, and obtains the compact result of k-means. Support Vector Machine (machine learning algorithm) is now applied on reduced feature space. This purposed K-SVM model gives higher accuracy with reduction in computation time.

T. Santhanam, M. S. Padmavathi in 2015 implemented [8] K-means along with Genetic Algorithm for dimensionality reduction and support vector machine to classify the data set. K-means algorithm is used to remove outliers and the noisy data. The optimal features are selected by using the genetic algorithm and then Support vector machine classifies the reduced data space using 10 fold cross validation technique. Genetic algorithm selects different features from original set of feature during each run. To obtain consistent results, the experiment was performed 50 times. The result shows that the purposed model achieves the accuracy of 98.82%.

A. Purwar, S.K.Singh in 2015 proposed [9] a prediction model for medical data with missing value imputation techniques, then analyzing these techniques by using K-means algorithm and choosing the best among them. Thus this model improves the quality of data by using the best imputation technique. Methods such as case deletion, most common method, concept most common, K-means clustering imputation, k-nearest neighbor etc. are applied to fill the missing data values in the data. The efficiency is calculated on three data sets namely Hepatitis, Wisconsin Breast Cancer and Pima Indians Diabetes from the UCI repository. This model achieved accuracy of 99.82% for Diabetes data set, 99.39% for Breast Cancer and 99.08% for Hepatitis data set. For Diabetes and Hepatitis data sets Concept Most Common (CMC) is chosen as the best method, and for Breast Cancer Case deletion is selected as best missing value imputation method.

A K Yadav, D Tomar, S Agarwal [10] in 2013 diagnosis of lung cancer. The lung cancer dataset is discussed with the domain experts and certain attributes with their impact factor are identified based on which the number of cluster is decided e.g. there is a possibility of cancer if the tumor size is greater than 3. On the basis on this clusters are formed. Then

the cluster will move left or right according to the impact of the next attribute on the cluster. The results show that the Foggy K-mean gives better result than the simple k-mean algorithm.

M.F.Akay in 2009 proposed [11] a system for breast cancer diagnosis by using support vector machine combined with the feature selection. Wisconsin breast cancer dataset from UCI repository is used for the experiment. This dataset contains nine features which are represented as a value between1 and 10. F-score for each feature is calculated and then sort the scores for the features. More discriminative feature has the larger F-score. Feature section is helpful in reducing the number of input feature in SVM classifier.

A Jain, A Rajavat, R Bhartiya, in 2012 proposed [12] modified k-mean clustering algorithm to cluster large datasets, the main motive is to find out the cluster centers which are very close to the final result for each iterative step. Modified k-mean clustering algorithm reduces problem of cluster error criterion and also avoids getting into locally optimal solution in some degree. They compare modified k-mean algorithm with k-mean clustering algorithm and the results shows that modified k-mean clustering algorithm take less time to execute than existing k-mean for small number of records as well as for large number of records. Modified k-mean algorithm is more strong to noise and outliers than K-means.

Poteras, C. M., Mihaescu, M. C.,&Mocanu, M in 2014 proposed [13] an optimized version of k-mean that reduces the problem of re-distribution of the data elements that will remain part of the same cluster during the next iteration. After a number of iterations only a few number of data elements change their cluster. While assigning the data element to the cluster there is no need to visit the entire data set, but just a small list of data objects. The implementation showed up to 70% reduction of the running time.

S Bharti, S. N Singh in 2015 stated [14] several algorithms like genetic algorithm, PSO, ANN that can be used in predicting heart disease. Combining these algorithms with the data mining techniques such as clustering, classification etc. or by combining these algorithms with one another will give better performance and accuracy.

IV. Challenges of Optimal Clustering

Cluster analysis divides **data** into groups (**clusters**) for the purposes of summarization or improved understanding. ... While **clustering** has a long history and a large number of **clustering** techniques have been developed in statistics, pattern recognition, **data mining**, and other fields, significant **challenges** still remain.

The potential problems with cluster analysis that we have identified in our survey are as follows: 1. The identification of distance measure: For numerical attributes, distance measures that can be used are standard equations like Euclidian,manhattan, and maximum distance measure. All the three are special cases of Minkowski distance .But identification of measure for categorical attributes is difficult.

2. The number of clusters: Identifying the number of clusters is a difficult task if the number of class labels is not known beforehand. A careful analysis of number of clusters is necessary to produce correct results. Else, it is found that heterogeneous tuples may merge or similarlytype's tuples may be broken into many. This could be catastrophic if the approach used is hierarchical.B'coz in hierarchical approach if a tuples gets wrongly merged in a cluster that action cannot be undone. While there is no perfect way to determine the number of Clusters, there are some statistics that can be analyzed to help in the process [15-16]. These are the Pseudo-F statistic, the Cubic Clustering Criterion(CCC), and the Approximate Overall R-Squared.

3. Lack of class labels: For real datasets (relational in nature as they have tuples and attributes) the distribution of data has to be done to understand where the class labels are?

4. Structure of database: Real life Data may not always contain clearly identifiable clusters. Also the order in which the tuples are arranged may affect the results when an algorithm is executed if the distance measure used is not perfect.With a structure less data(for e.g. Having lots of missing values), even

identification of appropriate number of clusters will not yield good results. For e.g. missing values can exist for variables, tuples and thirdly, randomly in attributes and tuples. If a record has all values missing, this is removed from dataset. If an attribute has missing values in all tuples then that attribute has to be removed described in [6]. A dataset may also have not much missing values in which case methods have been suggested in [17]. Also, three cluster-based algorithms to deal with missing values have been proposed based on the mean-and-mode method in [17].

5. Types of attributes in a database: The databases may not necessarily contain distinctively numerical or categorical attributes.They may also contain other types like nominal, ordinal, binary etc.So these attributes have to be converted to categorical type to make calculations simple.

6. Choosing the initial clusters: For partitonal approach, we find that most of the algorithms mention k initial clusters to be randomly chosen. A careful and comprehensive study of data is required for the same. Also, if the initial clusters are not properly chosen, then after a few iterations it is found that clusters may even be left empty. Although, a paper in [7] discusses a farthest heuristic based approach for calculation of centers.

V. Application of Optimal Clustering

Clustering analysis is broadly used in many applications such as market research, pattern recognition, data analysis, and image processing. Clustering can also help marketers discover distinct groups in their customer base. And they can characterize their customer groups based on the purchasing patterns.

Where clustering is been applied in various fields were some of the applications are:

Use of Clustering in Data Mining: Clustering is often one of the first steps in data mining analysis. It identifies groups of related records that can be used as a starting point for exploring further relationships. This technique supports the development of population segmentation models, such as demographic-based customer segmentation. Additional analyses using standard analytical and other data mining techniques can determine the characteristics of these segments with respect to some desired outcome. For example, the buying habits of multiple population segments might be compared to determine which segments to target for a new sales campaign.

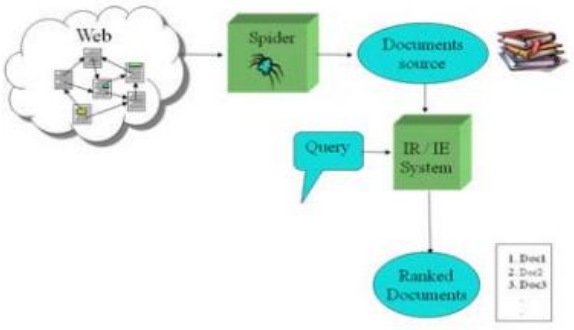
For example, a company that sale a variety of products may need to know about the sale of all of their products in order to check that what product is giving extensive sale and which is lacking. This is done by data mining techniques. But if the system clusters the products that are giving fewer sales then only the cluster of such products would have to be checked rather than comparing the sales value of all the products. This is actually to facilitate the mining process.

A. Application of Clustering in Text Mining

Text mining, also referred to as text data mining, roughly equivalent to text analytics, refers to the process of deriving high-quality information from text. High-quality information is typically derived through the devising of patterns and trends through means such as statistical pattern learning. Text mining usually involves the process of structuring the input text (usually parsing, along with the addition of some derived linguistic features and the removal of others, and subsequent insertion into a database), deriving patterns within the structured data, and finally evaluation and interpretation of the output. 'High quality' in text mining usually refers to some combination of relevance, novelty, and interestingness. Typical text mining tasks include text categorization, text clustering, concept/entity extraction, production of granular taxonomies, sentiment analysis, document summarization, and entity relation modeling Text mining consists of extraction information from hidden patterns in large text-data collectionsThe query is given in the system were the given query is been founded by using the search navigation system.Where the documents based on query search is been given here in the diagram.Where is been extracted using name extractor. From the authorization list the ranking details are viewed on it.

B. Working of Cluster in the Search Engines

Where information retrival system is works in the web documents on it. The document source is said to be the documents of the web page. The query is said to be the search engine. Using cluster the documents are classified based on the query in the information retrival system. The ranked documents represent the relevant details present in the documents which are relevant to the search of the query



It is the mining of the data in the web page…in the database websites.

C. Some other Applications of Clustering

Where the clustering is been used in Fields of applications on it.

- Data Mining
- Pattern recognition
- Image analysis • Bioinformatics
- Machine Learning
- Voice minim
- Image processing
- Text mining
- Web cluster engines
- Whether report analysis

VI. CONCLUSION

Data imbalance is a universal situation in practice, and cluster validation measures may not have the ability to capture its impact to K-means. Therefore, we have the following problems; divide dataset in various groups which are not known beforehand performance quality of clustering with reduced time finding the value of various clusters in dataset of K-means clustering. In this paper, a review of literatures on K-Means combine with genetic algorithm together is presented. In general, K-means has been widely studied in a great deal of research from both the optimization and the data perspectives. We have systematically analyzed 10 papers K-means with GA. It was observed from the literatures that many works have been done using GA to study number of groups not known in advance to cluster it and divide large-scale problem into several small problems to illustration very high performance solving small-scale combinatorial optimization. Thus, more work using GA is needed to investigate the K-means clustering enhancement in both measurement and determine the number of clusters. The outcomes from this literature are summarized as follow:

• It has been found that genetic algorithm has the ability to divide dataset in various numbers of groups which are not known beforehand. Employs label based representation and utilization of K-means techniques and strategies improves and enhances the offspring produced by the group based crossover and there is no need for fixing the cluster numbers. For the reduction of the problem complexity, Cluster processing used to get the center point.

• A large-scale problem is divided into several small problems and those methods show the GA is suitable for sub problems to improve speed and very high performance to solving small-scale combinatorial optimization.

REFRERENCES

[1]. Chen, Z.X. and Shixiong, "K-means Clustering Algorithm with improved Initial Center," in Second International Workshop on Knowledge Discovery and Data Mining, Moscow, 2009

[2]. Napoleon, D. and P.G. Lakshmi, "An Efficient K-means Clustering Algorithm for Reducing Time Complexity using Uniform Distribution Data Points," in Trends in Information Sciences and Computing (TISC), Chennai, 2010.

[3]. K.A.AbdulNazeer and M. P. Sebastian, “Improving the accuracy and efficiency of the kmeans clustering algorithm,” in International Conference on Data Mining and Knowledge

Engineering (ICDMKE), Proceedings of the World Congress on Engineering (WCE), Vol 1, July 2009, London, UK

[4]. Zhao, Weizhong, Huifang Ma, and Qing He. "Parallel k-means clustering based on map-reduce" In Cloud Computing, pp. 674-679. Springer Berlin Heidelberg, 2009

[5]. Fahim, A. M., A. M. Salem, F. A. Torkey, and M. A. Ramadan. "An efficient enhanced k-means clustering algorithm" Journal of Zhejiang University SCIENCE 7, No. 10, PP. 1626-1633, 2006

[6]. Yugal Kumar, Yugal Kumar, and G. Sahoo G. Sahoo. "A New Initialization Method to Originate Initial Cluster Centers for K-Means Algorithm" International Journal of Advanced Science and Technology 62, PP. 43-54, 2014

[7] Zheng, B., Yoon, S. W., & Lam, S. S. (2014). Breast cancer diagnosis based on feature extraction using a hybrid of K-means and support vector machine algorithms. Expert Systems with Applications, 41(4), 1476-1482.

[8] Santhanam, T., &Padmavathi, M. S. (2015). Application of K-Means and Genetic Algorithms for Dimension Reduction by Integrating SVM for Diabetes DiagnosisProcedia Computer Science, 47, 76-83

[9] Akay, M. F. (2009). Support vector machines combined with feature selection for breast cancer diagnosis. Expert systems with applications, 36(2), 3240-3247

[10] Purwar, A., & Singh, S. K. (2015). Hybrid prediction model with missing value imputation for medical data Expert Systems with Applications, 42(13), 5621-5631.

[11] Yadav, A. K., Tomar, D., &Agarwal, S. (2013, July). Clustering of lung cancer data using Foggy K-means In Recent Trends in Information Technology (ICRTIT), 2013 International Conference on (pp. 13-18)

[12] Jain, A., Rajavat, A., &Bhartiya, R. (2012, November). Design, Analysis and Implementation of Modified K-Mean Algorithm for Large Data-Set to Increase Scalability and Efficiency.In Computational Intelligence and Communication Networks (CICN), 2012 Fourth International Conference on (pp. 627-631).

[13] Poteras, C. M., Mihaescu, M. C., &Mocanu, M. (2014, September). An optimized version of the K-Means clustering algorithm In Computer Science and Information Systems (FedCSIS), 2014 Federated Conference on (pp. 695-699).

[14] Bharti, S., & Singh, S. N. (2015, May). Analytical study of heart disease prediction comparing with different algorithms In Computing, Communication & Automation (ICCCA), 2015 International Conference on (pp. 78-82)

[15] Milligan, G. W., & Cooper, M. C. (1985) "An Examination of Procedures for Determining the Number of Clusters in a Data Set," Psychometrika, 50, 159-179.

[16]Sarle, W. S., (1983) "Cubic Clustering Criterion," SAS Technical Report A-108, Cary, NC. SAS Institute Inc.

[17]. Fujikawa, Y. and Ho, T. (2002). Cluster-based algorithms for dealing with missing values.In Cheng, M.-S., Yu, P. S., and Liu, B., editors, Advances in Knowledge Discovery and Data Mining, Proceedings of the 6th Pacific-Asia Conference, PAKDD 2002, Taipei,Taiwan, volume 2336 of Lecture Notes in Computer Science, pages 549–554. New York:Springer.