ISSN: 2278 – 909X
International Journal of Advanced Research in Electronics and Communication Engineering (IJARECE)
Volume 7, Issue 5, May 2018

Performance enhancement for Text Data Mining using k means clustering based genetic optimization (KMGO)

Monika
Maharishi Dayanand University Rohtak

ABSTRACT

For discovering hidden patterns and structures in data, data mining can work better. Here we will study and comparison is done of on data mining techniques used for clusteringfrom Iris and Wine dataset with more accuracy as compare to existing work. K-Means is used for removing the noisy data and genetic algorithms for finding the optimal set of features of text mining. The experimental result proves that, the proposed model has attained an average accuracy.

Keywords: ***Data mining, Support Vector Machine, GA, K-means clustering.***

I. INTRODUCTION

Clustering techniques have become very popular in a number of areas, such as engineering, medicine, biology and data mining [1, 2]. A good survey on clustering algorithms can be found in [3]. The k-means algorithm [4] is one of the most widely used clustering algorithms. The algorithm partitions the data points (objects) into C groups (clusters), so as to minimize the sum of the (squared) distances between the data points and the center (mean) of the clusters. In spite of its simplicity, the k-means algorithm involves a very large number of nearest neighbor queries. The high time complexity of the k-means algorithm makes it impractical for use in the case of having a large number of points in the data set. Reducing the large number of nearest neighbor queries in the algorithm can accelerate it. In addition, the number of distance calculations increases exponentially with the increase of the dimensionality of the data [5-7]. Many algorithms have been proposed to accelerate the k-means. Elkan[10] suggests the use of triangle inequality to accelerate the k-means. In [11], it is suggested to use R-Trees. Nevertheless, R-Trees may not be appropriate for higher dimensional problems. Recently, Kernel - means [15] is an extension of the standard k-means algorithm that maps data points from the input space to a feature space through a nonlinear transformation and minimizes the clustering error in feature space. Thus, nonlinearly separated clusters in input space are obtained, overcoming the second limitation of k-means.

As seen in the literature, the researchers contributed only to accelerate the algorithm; there is no contribution in cluster refinement. In this study, we propose a new algorithm to improve the k-means using Genetic Algorithm (GA) is applied to refine the cluster to improve the quality.

II. PROBLEM IDENTIFICATION

Although K-mean algorithm eliminates the sensitivity to initial data for traditional k-means algorithm and obtains more stable and higher quality clusters, the number of categories must be determined in the traditional algorithm and K-Mean algorithm, randomly determined number of clusters often results in unsatisfied results. To solve this problem, a Genetic and K-means hybrid algorithm based on the optimized initial centers is proposed, which guarantees the convergence rate at the same time improving the clustering accuracy. Experiments will show that it is effective and feasible.

III. SYSTEM MODEL

The data mining supports the computational algorithms for analyzing the data automatically. Therefore the algorithms are implemented on data according to the nature of data and requirements of applications. Basically there different kinds of data mining techniques are available such as classification, association pattern mining, clustering and others. But in this work the key focus is placed on clustering based data analysis technique. The clustering is also termed in general discussion as grouping or categorization. Basically the clustering is an unsupervised manner of data analysis. In this technique the data features are compared each other for finding the suitable clusters. In other words the clustering techniques measure the internal similarity among the data objects for creating clusters.

In this context the similarity measurements between objects are computed either in terms of their similarity using the cosine similarity, and others or

using the dis -similarity by computing the distance between two data objects using Euclidean distance or others. But the key motive is to find how similar an object is, from other objects. In most of the time for computing the clusters the distance functions are used in linear manner but the linear distance is not much suitable for different applications. Thus in this presented work the Gaussian kernel is used for differentiating the data objects. In addition of that the some issues relevant to the clustering is also tried to improve such as optimization of centroid selection process for improving the fluctuating accuracy issue in k-means clustering.

IV. PROPOSED METHOD

A. K-Means Algorithm

The K-Means [11] is a simple and well known algorithm used for solving the clustering problem. The goal of the algorithm is to find the best partitioning of n objects into k clusters, so that the total distance between the cluster's members and its corresponding centroid, representative of the cluster is minimized. The algorithm uses an iterative refinement strategy using the following steps:

Steps 1: This step determines the starting cluster's centroids. A very common used strategy is to assign random k, different objects as being the centroids.

Steps 2: Assign each object to the cluster that has the closest centroid. In order to find the cluster with the most similar centroid, the algorithm must calculate the distance between all the objects and each centroid.

Steps 3: Recalculate the values of the centroids. The values of the centroid are updated by taking as the average of the values of the object's attributes that are part of the cluster.

Steps 4: Repeat Steps 2 and Step 3 iteratively until objects can no longer change clusters.

B. GO(Genetic optimization)

Genetic optimization [12] is stochastic methods for global search and optimization and belongs to the group of Evolutionary Algorithms. They simultaneously examine and manipulate a set of possible solution. Given a specific problem to solve, the input to GAs is an initial population of solutions called individuals or chromosomes. A gene is part of a chromosome, which is the smallest unit of genetic information. Every gene is able to assume different values called allele. All genes of an organism form a genome which affects the appearance of an organism called phenotype. The chromosomes are encoded using a chosen representation and each can be thought of as a point in the search space of candidate solutions. Each individual is assigned a score (fitness) value that allows assessing its quality. The members of the initial population may be randomly generated or by using sophisticated mechanisms by means of which an initial population of high quality chromosomes is produced. The reproduction operator selects (randomly or based on the individual's fitness) chromosomes from the population to be parents and enters them in a mating pool. Parent individuals are drawn from the mating pool and combined so that information is exchanged and passed to off-springs depending on the probability of the cross-over operator. The new population is then subjected to mutation and enters into an intermediate population. The mutation operator acts as an element of diversity into the population and is generally applied with a low probability to avoid disrupting cross-over results. Finally, a selection scheme is used to up- date the population giving rise to a new generation. The individuals from the set of solutions which is called population will evolve from generation to generation by repeated applications of an evaluation procedure that is based on genetic operators. Over many generations, the population becomes increasingly uniform until it ultimately converges to optimal or near-optimal solutions. The next section explains in detail the genetic algorithm used for the clustering problem. Algorithm 1 shows the various steps used in the proposed genetic algorithm.

Algorithm 1: Enhanced genetic algorithm with K-means.

Input: Problem P_0

Output: Solution $C_{final}(P_0)$

1) Begin

2) Generate initial population

3) Evaluate the fitness of each individual in the population

4) while (Not Convergence reached) do

5) Select individuals according to a scheme to reproduce;

6) Breed each selected pairs of individuals through crossover

7) Apply K-Means if necessary to each offspring according to Pk-Means

8) Evaluate the fitness of the intermediate population

9) Replace the parent population with a new generation

10) end.

1) Fitness function

The Notion of fitness is fundamental to the application of genetic algorithms. It is a numerical value that expresses the performance of an individual (solution) so that different individuals can be compared.

2) Representation

A representation is a mapping from the state space of possible solutions to a state of encoded solutions within a particular data structure. The encoding scheme used in this work is based on integer encoding. An individual or chromosome is represented using a vector of N positions, where N is the set of data objects. Each position corresponds to a particular data object, i.e. the i^{th} position (gene) represents the ith data object. Each gene has a value over the set {1,2,....k}. These values define the set of cluster labels.

3) Initial population

The initial population consists of individuals generated randomly in which each gene's allele is assigned randomly a label from the set of cluster labels.

4) Cross-over

The task of the cross-over operator is to reach regions of the search space with higher average quality. New solutions are created by combining pairs of individuals in the population and then applying a crossover operator to each chosen pair. The individuals are visited in random order. An unmatched individual i_l is matched randomly with an unmatched individual i_m. Thereafter, the two-point crossover opera- tor is applied using a cross-over probability to each matched pair of individuals. The two-point crossover selects two randomly points within a chromosome and then interchanges the two parent chromosomes be- tween these points to generate two new off- spring. Recombination can be defined as a process in which a set of configurations (solutions referred as parents) undergoes a transformation to create a set of configurations (referred as off-springs). The creation of these descendants involves the location and combinations of features extracted from the parents. The reason behind choosing the two point crossover is the results presented in [21] where the difference between the different crossovers is not significant when the problem to be solved is hard. In addition, the work conducted in [22] shows that the two-point crossover is more effective when the problem at hand is difficult to solve.

5)K-Means

By introducing local search at this stage, the search within promising areas is intensified. This local search should be able to quickly improve the quality of a solution produced by the crossover operator, with- out diversifying it into other areas of the search space. In the context of optimization, this rises a number of questions regarding how best to take advantage of both aspects of the whole algorithm. With regard to local search there are issues of which individuals will undergo local improvement and to what degree of intensity. However care should be made in order to balance the evolution component (exploration) against exploitation (local search component). Bearing this thought in mind, the strategy adopted in this regard is to let each chromosome go through a low rate intensity local improvement. The K-Means Algorithm described A is used for one iteration during which it seeks for a better clustering.

6) Mutation

The purpose of mutation which is the secondary search operator used in this work is to generate modified individuals by introducing new features in the population. By mutation, the alleles of the produced child individuals have a chance to be modified, which enables further exploration of the search space. The mutation operator takes a single parameter pm, which specifies the probability of performing a possible mutation. Let I = {c1, c2,…,ck } bean individual where each of whose gene ciis a cluster label. In our mutation operator, each gene c_i is mutated through flipping this gene's allele from the current cluster label c_i to a new randomly chosen cluster label if the probability test is passed. The mutation probability ensures that, theoretically, every region of the search space is explored. The mutation operator prevents the searching process from being trapped into local optima while adding to the diversity of the population and thereby in- creasing

the likelihood that the algorithm will generate individuals with better fitness values.

V. Result and discussion

Data files used in these experiments are chosen among a huge variety given by MATLAB. Here we are using genetic with three operators which are selection, crossover, mutation. In the taken data of genetic we assume the population size 100 and number of generation 1000.We are implementing both algorithms i.e. K-Mean and SVM using Genetic Algorithm in MAT lab and the corresponding results are shown below.

Dataset

The efficiency of the proposed KM-GO (K-mean genetic optimization) design is evaluated by conducting experiments on three datasets downloaded from UCI repository [1]. The description of the data sets used for evaluating the proposed method.

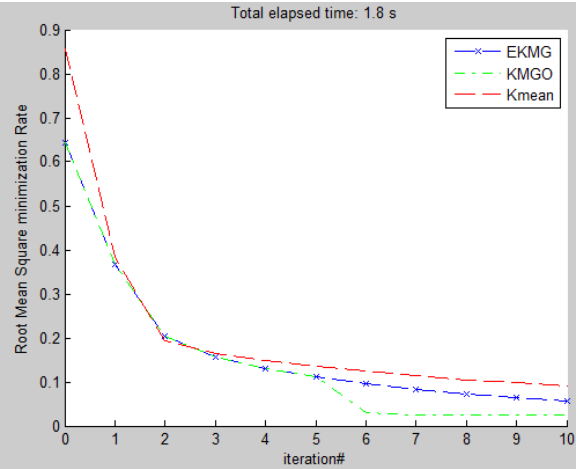

Figure 4: RMS value of EKMG, K-mean and KMGO on iteration

From Figure 4, the optimal threshold point value lies. For each method, we would like the RMS to be minimized. This will occur at the optimal iteration point value. By applying the optimal iteration point, we can produce the minimum RMS error and yield the optimal noise attenuation.

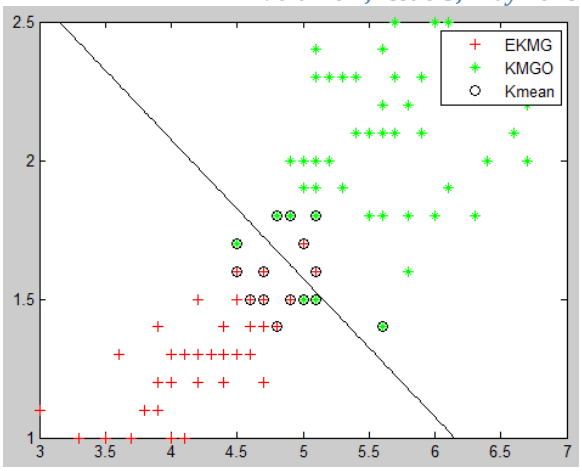

Figure 5: Hyper plane using SV-EKMG, EKMG and K-mean

Above figure shows that the data classified using K-mean and Genetic algorithm get best solution as seen in hyper plane.

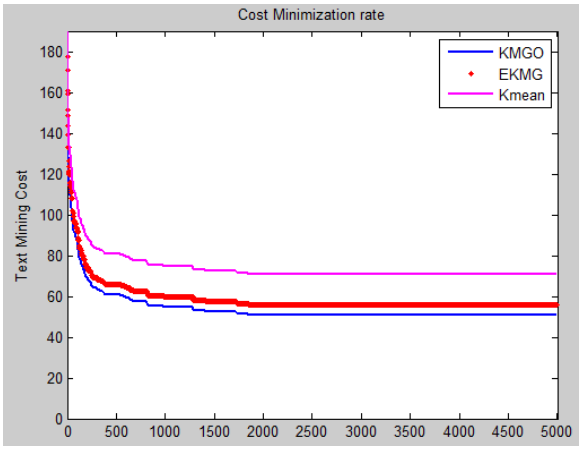

Figure 6: Cost minimization rate for Text mining cost between KMGO, EKMG and K-mean.

The Text mining cost value versus generation number curves of the Text mining cost, cost mininmization rate and KMGO value optimizations for the head-and-neck case. Figure 6 shows the best Text mining costafter the final generation for different crossover rate ranging from 0 to 5000.

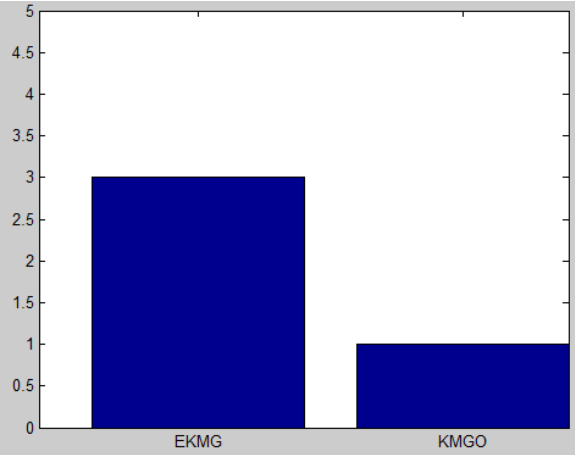

Figure 7: Accuracy of two different data for EKMG (Existing) and KMGO (Proposed)

In order to evaluate a clustering method, it is necessary to define a measure of agreement between two partitions of same data sets. The first experiment is carried out with the Iris problem. This problem is considered as main data mining clustering problem. The Wine problem from UCI repository is considered as second experimental. We can see that the accuracy of IRIS is more than Wine for KMGO.

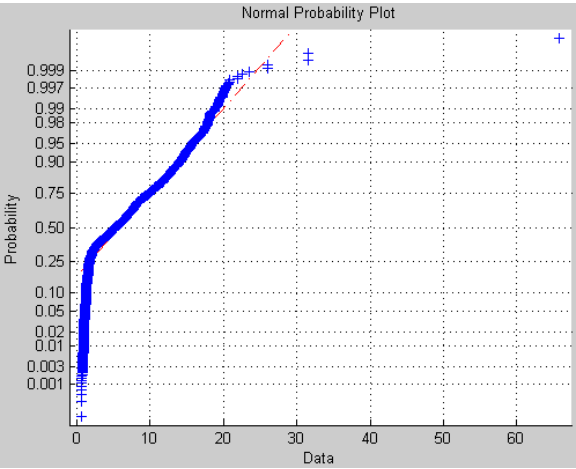

Figure 8: probability plot of proposed method

VI. Conclusion

This paper includes algorithms like K-means algorithm for clustering, and GO (Genetic optimization) for finding the optimal data. From the observation Existing possess the least classification and clustering accuracy and Proposed provide the better classification clustering accuracy results.

REFERENCES

[1] Lv T., Huang S., Zhang X., and Wang Z, Combining Multiple Clustering Methods Based on Core Group. Proceedings of the Second International Conference on Semantics, Knowledge and Grid (SKG'06), pp: 29-29, 2006.

[2] Nock R., and Nielsen F., On Weighting Clustering. IEEE Transactions and Pattern Analysis and Machine Intelligence, 28(8): 1223-1235, 2006.

[3] Xu R., and Wunsch D., Survey of clustering algorithms. IEEE Trans. Neural Networks, 16 (3): 645-678, 2005.

[4] MacQueen J., Some methods for classification and analysis of multivariate observations. Proc. 5th Berkeley Symp.Math. Stat. and Prob, pp: 281-97, 1967.

[5] Kanungo T., Mount D.M., Netanyahu N., Piatko C., Silverman R., and Wu A.Y., An efficient k-means clustering algorithm: Analysis and implementation. IEEE Trans. Pattern Analysis and Machine Intelligence, 24 (7): 881-892, 2002.

[6] Pelleg D., and Moore A., Accelerating exact k means algorithm with geometric reasoning. Proceedings of the fifth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, New York, pp. 727-734, 1999.

[7] Sproull R., Refinements to Nearest-Neighbor Searching in K-Dimensional Trees. Algorithmica, 6: 579-589, 1991.

[8] Bentley J., Multidimensional Binary Search Trees Used for Associative Searching. Commun. ACM, 18 (9): 509-517, 1975.

[9] Friedman J., Bentley J., and Finkel R., An Algorithm for Finding Best Matches in Logarithmic Expected Time. ACM Trans. Math. Soft. 3 (2): 209-226, 1977.

[10] Elkan, C., Using the Triangle Inequality to Accelerate k-Means. Proceedings of the Twentieth International Conference on Machine Learning (ICML-2003), pp. 609-616, 2003.

[11] J. B. MacQueen, "Some methods for classification and analysis of multivariate observation," in *Proc. Berkeley Symposium on Mathematical Statistics and Probability*, Univ. of California Press, 1967.

[12] D. E. Goldberg, *Genetic Algorithms in Search, Optimization, and Machine Learning*, Addison-Wesley, New York, 1989.