RESEARCH ARTICLE OPEN ACCESS

A High Dimensional Clustering Scheme for Data Classification

Navjot Gwal* and Tejalal choudhary**
*(Research Scholar, Department of Computer Science & Engineering, Sushila Devi Bansal College of Technology, Indore, Madhya Pradesh, INDIA
navjot.gwal18@gmail.com)
**(Assistant Professor, Department of Computer Science & Engineering, Sushila Devi Bansal College of Technology, Indore, Madhya Pradesh, INDIA
tejalal.choudhary@gmail.com)

ABSTRACT
The data mining is the knowledge extraction or finding the hidden patterns from large data these data may be in different form as well from different resources. The data mining systems can be used in various research domains like health, share market analysis, super market, weather forecasting and many other domains. Data mining systems use the computer oriented algorithms. These algorithms can be categorized as supervised and unsupervised respectively. The classification or prediction algorithms belong to supervised category and clustering algorithms are the type of unsupervised. The clustering is an approach to group the similar data objects to the same cluster and the main aspect of clustering is that the distance between data objects in same cluster should be as minimum as and distance of objects in inter cluster should be high. k-means is one of the most common clustering algorithm. K-means is very easy to use and efficient but has also some weakness because of random or inappropriate selection of initial centroids so need to improve k-means. The proposed work is an attempt to improve k means by using genetic algorithm for selection of initial cluster centroid.

Keywords: Clustering, data mining, genetic algorithm, k-means clustering and unsupervised learning.

I. INTRODUCTION

In data mining [1] different techniques are utilized to analyse data using computer based algorithms. These algorithms measure the similarity and dissimilarity among available set of data. Using this evaluation the data patterns are recovered. The analysis of data is performed in both supervised and unsupervised manner. The supervised technique supports the classification techniques and the unsupervised technique supports the clustering [2][3] techniques [4][5] for data analysis. The proposed work is devoted to understand and develop an efficient and accurate unsupervised technique which provides efficient results and can resolve the deficiencies of available clustering techniques.

In order to develop such an efficient and accurate clustering algorithm a number of clustering algorithms such as K-means clustering, C-means clustering and other techniques are studied. Among them the k-means algorithm found more effective and frequently used algorithm. Therefore some key issues are observed and targeted to improve. Therefore the k-means algorithm has been investigated in detail and the recovered facts shows that k-means has fluctuating accuracy, high error rate and others weakness. In order to resolve the issues in traditional k-means [6] algorithm a new algorithm is proposed that uses a well known algorithm called genetic algorithm [7] to find the optimum solution to search problems.

In this presented work the k-means algorithm is modified for finding the better and stable performance of clustering. The improved K-means overcomes the problems present in the traditional k-means algorithm by pre-processing the datasets. The improved k-means algorithm is a single pass algorithm. The improvement is made on the traditional clustering approach by implementing genetic algorithm based cluster centroids selection. If the optimum centroids from the data are selected then the issues such as accuracy fluctuation and performance issues are minimized. Therefore the genetic algorithm is implemented for selection of initial centroids.

II. BACKGROUND

Liang Bai et al [9] proposed a new clustering technique based on fuzzy for categorical data. This work added the information about between clusters to the k-mode objective function that reduces the dispersion in the cluster and separates the clusters more efficiently. By using well suited update formulas to get the local optimum solution about to this modified k-mode objective function and ultimately generate cluster prototype.

Nenad Tomasev et al [10] focus to provide an advanced clustering technique for high dimensional data. High dimensional data have the feature about to lower dimension subspace that irritate to mining system. But this proposed technique don’t concentrate such dark issue instead uses high dimension phenomena exploited data points that occur frequently in k-nearest neighbor list for clustering. This technique generates the clusters of hyper spherical shapes and need to improve to get arbitrary shape cluster.

Mitchell Yuwono et al's [11] the base for this research work is particle swam clustering and proposed a new approach named Rapid centroid estimation that increase the efficiency of the particle trajectories causes higher deduction in computation complexity and improve the particle swarm clustering update rule. The experimental results showed that time consumed in iterative process in Rapid centroid estimation is very less than particle swarm clustering as well as modified swarm clustering.

Jian Zhang et al [12] to cluster the data taking the advantage of shadowed sets and well known clustering particle swarm optimization and proposed the new modified version of fuzzy c-means. Merges the search capability of particle swarm optimization and vagueness balance capability of shadowed sets and control the cluster overlapping problem. These algorithm gives a new concept of optimally find the cluster number automatically. This work increases the effectiveness of clustering

.

Anirban Mukhopadhyay et al [13] uses the multi-objective evolutionary algorithms for data mining. Provided two new techniques that use the multi-objective evolutionary algorithm, the first technique for feature selection in classification and second for feature selection in association rule mining clustering and other data mining techniques. This research work scopes the multi-objective evolutionary algorithms in future data mining researches.

III. PROPOSED WORK

Figure 1 shows the simulation methodology of the system that describes the effectiveness of the genetic algorithm based clustering approach.

In this given system as shown in figure 1 first provide the input data to system. Then data is prepossessed after that knowledge or patterns are extracted from the data. In next step user select the appropriate algorithm by which the clustering operation is done by system. Here two different algorithms namely traditional k-means clustering and proposed algorithm is implemented. The user selected algorithm consumes the data and produces the clusters of data. During this the data is grouped into the given number of clusters according to the selected algorithm. The prepared model is evaluated for finding the performance of algorithm. Then the validation process of data model is performed with random set of test data.

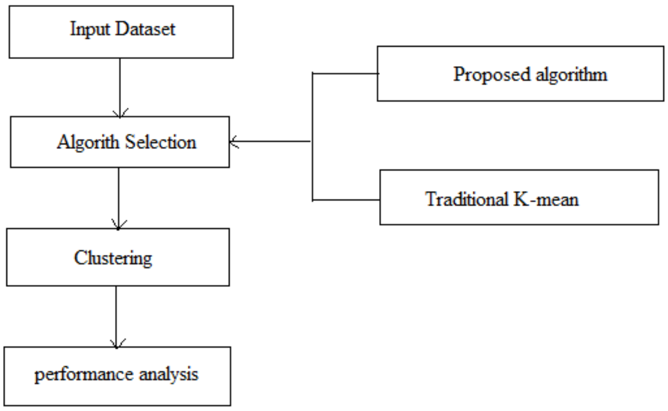

Figure 1. Simulation system

A. Algorithm study

This section provides the detailed study about the traditional algorithms that are used for implementation of proposed model.

K-means algorithm

The K-Means clustering algorithm is a partition-based cluster analysis method [6]. According to the algorithm firstly select k objects as initial cluster centroids, then calculate the distance between each object and each cluster centroids and assign it to the nearest cluster, update the averages of all clusters, repeat this process until the criterion function converged. Square error criterion for clustering

$$E = \sum_{i=1}^{k}\sum_{j=1}^{n_i}\left\|x_{ij} - m_i\right\|^2$$

x_{ij} is the sample j of i-class, $m_i$is the centroids of i-class, n_i i is the number of samples of i-class. K-means clustering algorithm is simply described step by step by algorithm 1.

Algorithm 1: The K-means clustering algorithm

Input: N objects to be cluster (xj, Xz … xn), the number of desired clusters k.

Output: k clusters of N input data objects.

Process:

1. Arbitrarily select k objects as initial cluster centroids $(m_1, m_2, \dots, m_k)$;

2. Calculate the distance between each object Xi and each cluster centroid, then assign each object to the nearest cluster, formula for calculating distance as:

$$d(x_i, m_i) = \sqrt{\sum_{j=1}^{d}\left(x_i - m_{j1}\right)^2}, i = 1 \dots N, j = 1 \dots k$$

$d(x_i, m_i)$ is the distance between data i and cluster j.

3. Calculate the mean of objects in each cluster as the new cluster centroids,

$$m_i = \frac{1}{N}\sum_{j-1}^{n_i} x_{ij}\ , i = 1,2,\ldots,K$$

N_i is the number of samples of current cluster i;
4. Repeat 2) 3) until the criterion function E converged, return $(m_1, m_2, \ldots, m_k)$ Algorithm terminates.

Genetic algorithm

Genetic algorithm is genetically inspired search process that finds the optimum solution in huge search space. The genetic algorithms use the three main concepts for solution discovery: reproduction, natural selection and diversity of the genes [8]. The brief description of the overall search process is given ahead.

Generate initial population: Initially the genetic algorithms are initiated with the randomly generated sequences, with the allowed alphabets for the genes.

Selection: This is a process of selecting the optimum symbols among all individuals, the scaling of sequences is performed and using these best n individuals are transferred to the new generation.

Crossover: The crossover is basically the process of recombination; the individuals are chosen by selection and recombined with each other.

Mutation: The random change on some of the genes guarantees, it is still possible to generate them using the mutation process by randomizing the search.

New generation: The selected individuals from the selection process combined with those genes that are processed with the crossover and mutation for next generation development. [9].

Check for termination of the algorithm: This is the controlling approach of genetic algorithm. It is possible to stop the genetic optimization iterative process by using either value of the fitness function or Maximum number of iterations and or fixing the number of generations.
Complete description of the traditional genetic algorithm is specified in algorithm 2.

Algorithm 2: Traditional Genetic Algorithm

```
Input:
        Instance $\Pi$,
        Size of population $\alpha$,
        Rate of elitism $\beta$,
        Rate of mutation $\gamma$,
        Number of iterations $\delta$

// initialization
    1.  Generate $\alpha$ feasible solutions randomly;
    2.  Save them in the population $P$OP;

//Loop until the terminal condition
    3.  For $i$ = 1 to $\delta$ do

//Elitism based selection
    4.  Number of elitism $n$e = $\alpha \cdot \beta$;
    5.  Select the best $n$e solutions in $P$o$P$ and save
        them in $P$o$P_1$;

//Crossover
    6.  Number of crossover $n$c = $(\alpha - n$e$)/2$ ;
    7.  For $j$ = 1 to $n$c do
            a.  Randomly select two solutions $X_A$
                and $X_B$ from $P$o$P$ ;
            b.  Generate $X_C$ and $X_D$ by one-point
                crossover to $X_A$ and $X_B$;
            c.  Save $X_C$ and $X_D$ to $P$o$P_2$ ;
    8.  End for

//Mutation
    9.  For $j$ = 1 to $n$c do
            a.  Select a solution $X_j$ from $P$o$P_2$;
            b.  Mutate each bit of $X_j$ under the rate
                $\gamma$ and generate a new solution $X_j'$ ;
            c.  If $X_j'$ is unfeasible
                    i.  Update $X_j'$ with a feasible
                        solution by repairing $X_j'$ ;
            d.  End if
            e.  Update $X_j$ with $X_j'$ in $P$o$P_2$;
    10. End for

//Updating
    11. Update $P$o$P$ = $P$o$P_1$ + $P$o$P_2$ ;
    12. End for
    13. Returning the best solutions
    14. Return the best solution $X$ in $P$o$P$ ;
```

Proposed algorithm

This section demonstrates the designing of the proposed algorithm. Thus utilization of traditional algorithms is performed for achieving the enhanced algorithm processing. The summarized steps of data analysis are given in algorithm 3.

Algorithm 3: Proposed Algorithm

Input: Dataset, elitism rate, population size, number of clusters

Output: Clustered data
Process:
1. Process input dataset
2. Find unique attributes from available set of data in dataset
3. Using the obtained attributes generate population
4. Evaluated population using genetic algorithm

```
5. If number of sequences = Number of clusters
        Stop genetic algorithm
6. Else if number of evaluation cycles = Number of iterations
        Stop genetic algorithm
7. End if
8. Select most fit sequences as cluster centroids
9. Apply K-means with selected centroid
10. Return clustered data
```

The algorithm 3 shows the proposed algorithm for clustering, in this algorithm the input data is processed to make clean. After pre-processing of data the attribute wise data analysis is performed thus first for each attributes the number of attributes is extracted then using these unique symbols the population is generated. The newly generated population is similar in size as the input dataset. Now the remaining steps namely selection, cross over and mutation is performed on the generated population. The genetic algorithm works till entire data is evaluated. There are two termination conditions are used first the number of clusters and number of sequences are similar (this returns most optimum tree points in data that are in similar in distance). Or second the numbers of iterations are reached, thus the fit values as centroid is selected. Now the k-means clustering algorithm is used to perform clustering using the selected centroids. Thus process guarantees to provide effective clustering.

IV. RESULT ANALYSIS

The implementation of the desired technique completed and that is required to evaluate the implemented model. Thus this chapter contains the evaluated parameters and comparative results analysis of the proposed system.

A. Accuracy

In data mining and machine learning applications the amount of input samples are correctly recognized is known as the accuracy of the classifier or algorithm. The accuracy can be estimated using the given formula.

$$Accuracy = \frac{Total\ correctly\ identified\ samples}{Total\ samples\ to\ classify} X100$$

The comparative accuracy of the proposed clustering algorithm is given using figure 2, in this diagram the proposed clustering algorithm is demonstrated using blue line and the traditional K-means clustering is given by the red line. For reporting performance X axis contains the dataset size and the Y axis contains the percentage accuracy. The results demonstrate the proposed algorithm provide high accurate results as compared to the traditional K-means algorithm.

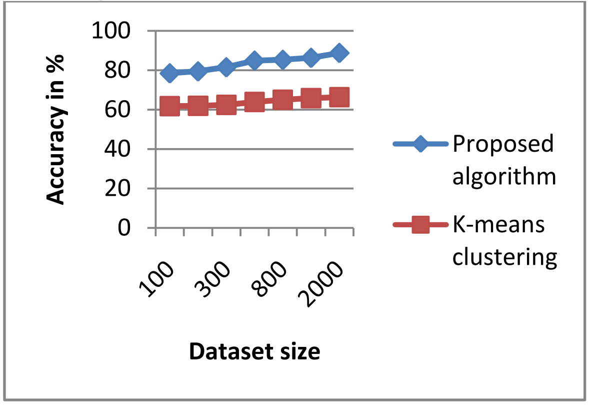

Figure 2. Accuracy

B. Error rate

The error rate of the algorithm demonstrates the amount of data which is not correctly identified during classification. The error rate of an algorithm can be evaluated using the below given formula.

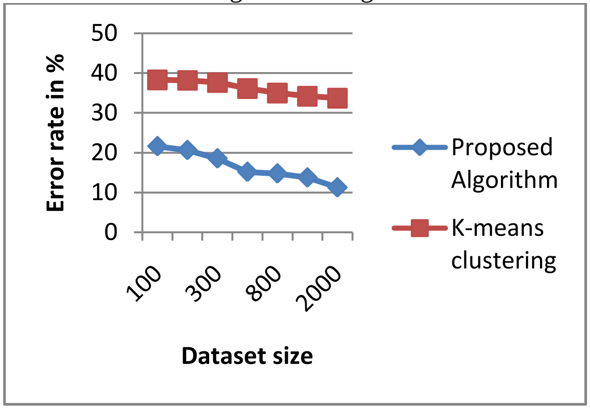

Figure 3. Error rate

$$Error\ rate = \frac{Total\ incorrectly\ identified\ samples}{Total\ samples\ as\ input} X100$$

Or

$$Error\ rate = 100 - Accuracy$$

The given figure 3 shows the comparative percentage error rate observed during evaluation of the implemented algorithms. To represent the performance of the system the X axis simulates the different data set size used for experiments with the system and the Y axis shows the percentage error rate. On the basis of experimental results the performance of the proposed algorithm improved and produces less error as compared to the traditional algorithms, additionally decreasing the error rate of shows improving performance of algorithm.

C. Memory used

The amount of main memory required to successfully execute the algorithm is known as the memory consumption or space complexity of the algorithm. The amount of memory consumed during different experiments is reported using figure 4.

In order to show the performance of algorithm X axis of the diagram shows the size of dataset used with experiments and the Y axis shows the amount of main memory consumed during experimentations in terms of KB (kilobytes). On the basis of experimental results the performance of the proposed algorithm in terms of memory usage is higher as compared to traditional algorithms. In other words the memory requirement of algorithms increases as the amount of data for experiments increases. The excessive memory used by the system shows the overhead of the encryption and decryption of the data during the clustering of data.

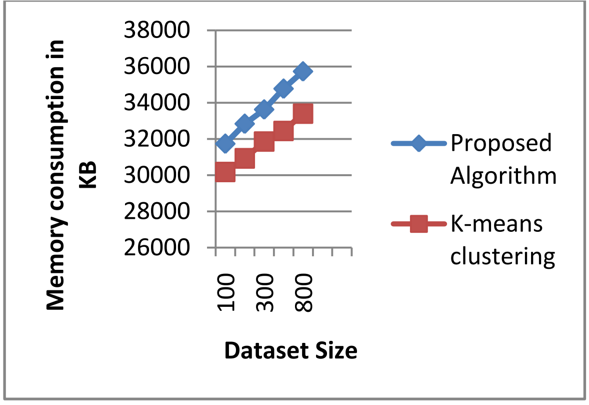

Figure 4. Memory consumption

D. Time Complexity

The amount of time required to perform clustering by algorithm is known as the time complexity.

The time complexity of the proposed technique and K-means clustering algorithm is given in figure 5. In this diagram the X axis shows the dataset size and the Y axis shows the time complexity of the algorithm in terms of MS (milliseconds). According to the given diagram the proposed algorithm consumes more time as compared to the traditional algorithm thus the proposed algorithm is complex as compared to traditional method of clustering.

V. CONCLUSION AND FUTURE WORK

The proposed clustering algorithm is hybrid technique of the unsupervised learning. In this technique cluster heads selection or centroid selection is based on the genetic algorithm. Thus in first step the genetic algorithm is utilized for centroid selection and in next step clustering is performed on data. In order to show effectiveness of the clustering algorithm the proposed clustering technique is compared with the traditional K-means clustering algorithm.

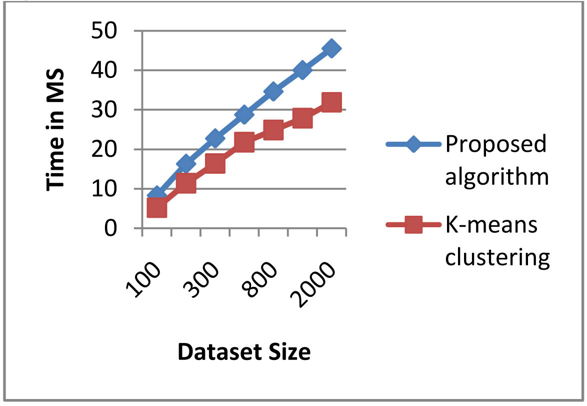

Figure 5. Time complexity

The implementation of the proposed work is performed using JAVA technology and its performance is evaluated. The performance of the algorithm is measured in terms of accuracy, error rate, memory usage, and time complexity. After evaluation of the proposed technique the performance of algorithm is compared with the K-means clustering algorithm. On the basis of different set of experiments the performance of algorithms are summarized using table 1.

Table 1. Performance summary

S. No.	Parameters	Proposed algorithm	K-means clustering
1	Accuracy	High	Low
2	Error rate	Low	High
3	Memory usage	High	Low
4	Time consumption	High	Low

Evaluated results show the performance of the proposed clustering algorithm is found optimum and less fluctuating as compared to the traditional clustering algorithm. Thus the proposed clustering algorithm is comparatively more adoptable than traditional clustering algorithms.

The key aim of designing an enhanced clustering algorithm is completed yet and that found using the effective centroid selection the performance of clustering algorithm is enhanced. Thus the proposed clustering technique is accurate most of the time; the only limitation is that it consumes more time and memory as compared to the traditional clustering algorithm. Thus that is required to collect more literature for enhancing the performance of the clustering algorithm.

REFRENCES

[1] Data mining Concepts and Techniques, Second Edition, Jiawei Han and Micheline Kamber, http://akademik.maltepe.edu.tr/~kadirerdem/772s_Data.Mining.Concepts.and. Techniques.2nd.Ed.pdf.

[2] Data Clustering: A Review, A.K. JAIN, M.N. MURTY, P.J. FLYNN, © 2000 ACM 0360-0300/99/0900–0001.

[3] Raza Ali, Usman Ghani, Aasim Saeed, "Data Clustering and Its Applications". http://members.tripod.com/asim_saeed/paper.htm

[4] A Comparative Study of Data Clustering Techniques, Khaled Hammouda, Department of Systems Design Engineering, University of Waterloo, Waterloo, Ontario, Canada N2L 3G1.

[5] Khushbu Patel, "Analysis of Various Database Using Clustering Techniques", Volume-3, Issue-7, July-2014 • ISSN No 2277 – 8160.

[6] An improved K-Means clustering algorithm, Juntao Wang, Xiaolong Su, 978-1-61284-486-2/111$26.00 ©2011 IEEE.

[7] Genetic Algorithms for Optimization, Programs for MATLAB ® Version 1.0 User Manual.

[8] Jifeng Xuan, He Jiang, ZhileiRen, "Pseudo Code of Genetic Algorithm and Multi-Start Strategy Based Simulated Annealing Algorithm for Large Scale Next Release Problem", Dalian University of Technology.

[9] Liang Bai, Jiye Liang, Chuangyin Dang, Fuyuan Cao, "A novel fuzzy clustering algorithm with between-clusterinformation for categorical data", Fuzzy Setsand Systems (2012), 2012 Elsevier B V All rights reserved.

[10] Nenad Tomasev, Milos Radovanovi, Dunja Mladeni, and Mirjana Ivanovi, "The Role of Hubness in Clustering High-Dimensional Data", IEEE Transactions on Knowledge and Data Engineering, Revised January 2013.

[11] Mitchell Yuwono, Steven W. Su, Bruce D. Moulton, and Hung T. Nguyen, "Data Clustering Using Variants of Rapid Centroid Estimation", Copyright 2012 IEEE. Personal use of this material is permitted.

[12] Jian Zhang and Ling Shen, "An Improved Fuzzy c-Means Clustering Algorithm Based on Shadowed Sets and PSO", Hindawi Publishing Corporation Computational Intelligence and Neuroscience Volume 2014, Article ID 368628, 10 pages.

[13] Anirban Mukhopadhyay, Ujjwal Maulik, Sanghamitra Bandyopadhyay, and Carlos A. CoelloCoello, "Survey of Multi-Objective Evolutionary Algorithmsfor Data Mining: Part-II", IEEE Transactions on Evolutionary Computation · February 2014.