International Journal of Computer Theory and Engineering, Vol.3, No.1, February, 2011
1793-8201

Mining Clusters in Data Sets of Data Mining: An Effective Algorithm

Singh Vijendra ,Sahoo Laxman and Kelkar Ashwini

***Abstract*—This paper propose a new clustering algorithm (GACR) based on genetic algorithm. The searching capability of genetic algorithms is exploited in order to search for appropriate cluster centers in the feature space such that a similarity metric of the resulting clusters is optimized. The chromosomes, which are represented as strings of real numbers, encode the centers of a fixed number of clusters. A chromosome reorganization method is proposed, which may effectively remove the degeneracy for purpose of more efficient search. A new crossover operator that exploits a measure of similarity between chromosomes is also presented. Adaptive probabilities of crossover and mutation are employed to prevent the convergence of the GA to a local optimum. The features of this algorithm are investigated and the performance is evaluated experimentally using real and synthetic datasets with K-means and GCA [10].The experimental result demonstrates that the GACR clustering algorithm has high performance, effectiveness and flexibility.**

***Index Terms*—Adaptive probabilities, Clustering, Evolutionary computation, Genetic algorithm**

I. Introduction

Data mining is a process for discovering previously unknown potentially useful and under-stand able patterns from large amounts of data [1]. The novelty and the comprehensibility of the mining results and the scalability of the algorithm are all indispensable for the success of a data-mining project. All data mining tasks can be categorized into two types: supervised tasks and unsupervised tasks. Supervised data-mining tasks have datasets that contain both the explanatory variables and the dependent variables, and the objective is to discover the relationships between the explanatory variables and the dependent variables. On the other hand, unsupervised mining tasks have datasets that contain only the explanatory variables, with the objective to explore and generate hypotheses about the hidden structures of the data. Clustering is one of the most common unsupervised data mining methods that explore the hidden structures embedded in a dataset. Clustering is the process of grouping a set of objects into clusters so that objects within a cluster are similar to each other but are dissimilar to objects in other clusters. Clustering has been effectively applied in a variety of engineering and scientific disciplines such as psychology, biology, medicine, computer vision, communications, and remote sensing [2]. Cluster analysis organizes data (a set of patterns, each pattern could be a vector measurements) by abstracting underlying structure. The grouping is done such that patterns within a group (cluster) are more similar to each other than patterns belonging to different groups. Thus, organization of data using cluster analysis employs some dissimilarity measure among the set of patterns. The dissimilarity measure is defined based on the data under analysis and the purpose of the analysis. Various types of clustering algorithms have been proposed to suit different requirements.

Traditional clustering algorithms primarily distinguish between hierarchical and partition [2], [3].Based on the clustering criterion adopted by the algorithm [4]. We can categorize existing clustering algorithms into the following classes.

The first class employs a local concept of clustering based on the idea that neighboring data points should share the same cluster. Algorithms implementing this principle are density-based methods [5], [9], nearest neighbor methods [6] and methods like single link agglomerative clustering [1]. These methods are well suited to detect clusters of arbitrary shapes; however, they are not robust when there is little spatial separation between the clusters.

The second class methods are generally implemented by keeping intra clusters variation (i.e., variation between same-cluster data points or between data points and cluster representatives) small. This category includes algorithms like K-means [2],[3], average link agglomerative clustering [1], and model-based clustering approaches [7], [8]. These methods tend to be very effective for spherical and well-separated clusters, but they may fail for more complicated cluster structures.

The third class performs simultaneous row column clustering. Typical examples of this kind are bi-clustering algorithms [11]–[13].The goal of these techniques is to identify subgroups of rows and subgroups of columns, by performing simultaneous clustering of both rows and columns of the data matrix, instead of clustering these two dimensions separately. Therefore, bi-clustering techniques produce local models, whereas clustering approaches compute global models. The clusters identified by these algorithms are not mutually exclusive or exhaustive. A data point may belong to no cluster or one or more clusters.

Manuscript received on December 20, 2009.

Vijendra Singh is with Faculty of Engineering and Technology,Mody Institute of Technology and Science ,Lakshmangarh, Sikar, Rajasthan, India(phone: 919829668880; email: d_vijendrasingh@yahoo.co.in).

Laxman Sahoo is with NIEC ,Luck now, UP, India(email: laxmansahoo@yahoo.com.).

Ashwini Kelkar is with Faculty of Engineering and Technology,Mody Institute of Technology and Science ,Lakshmangarh, Sikar, Rajasthan, India.

The fourth optimizes several validity measures that can capture the different characteristics of the data set. This kind of algorithms can be subdivided into clustering ensembles [14]–[16], which combine the resulting solutions into a single one with better quality, and multi objective clustering methods [11],[17], which provide an estimate of the quality of all individual clustering solutions and determine a set of potentially promising clustering solutions. Though often more robust and yield higher quality results than individual clustering methods, these algorithms are not without drawbacks. Clustering ensembles operate with homogenous objective functions. Therefore, good clustering results may become diluted by weak results in an ensemble. For the multi objective approaches, the construction of the promising solutions set is a difficult conceptual problem, since clustering algorithms often are not accompanied by a measure of the goodness of the detected clusters. Each clustering result should be judged not only by the clustering algorithm that generated it, but also by external assessment criteria.

The K-means algorithm is one of the more widely used algorithms. However, it is well known that K-means algorithm is sensitive to the initial cluster centers [3] and easy to get stuck at the local optimal solutions [18]. Moreover, when the number of data points is large, it takes enormous time to find the global optimal solution [19]. In order to improve the performance of the K-means algorithm, a variety of methods have been proposed [20]–[21].

In order to overcome the limitations of traditional clustering algorithms, some attempts have been made to use genetic algorithms for clustering data sets. Genetic algorithms (GAs) [22] are randomized search and optimization techniques guided by the principles of evolution and natural genetics. They are efficient, adaptive and robust search processes, performing multi-dimensional search in order to provide near optimal solutions of an evaluation (fitness) function in an optimization problem. Since the problem of clustering may be viewed as searching for a number of clusters in the feature space such that a given clustering metric is optimized, application of GAs to this problem seems natural and appropriate.Murthy and Chowdhury [24] have considered a partition to be encoded as a string of length n, where n is the number of data points. The i_{th} element of the chromosome represents the cluster number to which the corresponding point belongs. A comparison of the performance of their algorithm, subsequently referred to as the GA-clustering algorithm, with that of the K-means algorithm is provided in [24].Tseng and Yang [25] proposed a genetic algorithm based approach for the clustering problem. Their method consists of two stages, nearest neighbor clustering and genetic optimization. Bandyopadhyay and Maulik[26] applied the variable string length genetic algorithm with the real encoding of the coordinates of the cluster centers in the chromosome to the clustering problem. A K-mean clustering is performed taking the membership degrees and prototype locations as the parameters for the GA in [15].Lai [27] adopted the hierarchical genetic algorithm to solve the clustering problem. In the proposed method, the chromosome consists of two types of genes, control genes and parametric genes. The control genes are coded as binary digits. The total number of 1 represents the number of clusters. The parametric genes are coded as real numbers to represent the coordinates of the cluster centers. The relationship between the control genes and the parametric genes is that the activation of the latter is governed by the value of the former. If the value of a control gene is 1, then the associated parametric genes due to that particular active control gene are activated; otherwise the associated parametric genes are disabled. Lin [28] presented a genetic clustering algorithm based on a binary chromosome representation. The proposed method selects the cluster centers directly from the data set. With the aid of a look-up table, the distances between all pairs of objects are saved in advance and evaluated only once throughout the evolution process.Bandyopadhyay and Saha [29] proposed an evolutionary clustering technique that uses a new point symmetry-based distance measure. The Kd-tree based nearest neighbor search is used to reduce the complexity of finding the closest symmetric point. Adaptive mutation and crossover probabilities are used. The proposed GA with point symmetry distance based clustering algorithm is able to detect any type of clusters, irrespective of their geometrical shape and overlapping nature, as long as they possess the characteristic of symmetry. Singh and Sahoo [10] proposed a genetic clustering algorithm (GCA) that finds a globally optimal partition of a given data sets into a specified number of clusters. The algorithm is distance-based and creates centroids.

In this paper, a genetic algorithm with chromosome reorganize (GACR) is introduced to enhance the performance of clustering. In GACR, the degeneracy of chromosome is effectively removed, which makes the evolution process converge fast. Furthermore, a new crossover operator is introduced.

This paper is organized as follows. Clustering problem is briefly described in section II. In section III Genetic Algorithm with Chromosome Reorganization (GACR) is explained. Section IV contains data description and result analysis. Finally, we conclude in Section V.

II. Clustering Problem

Clustering is a formal study of algorithms and methods for classifying objects without category labels. A cluster is a set of objects that are alike, and objects from different clusters are not like. The set of n objects $X=\{x_1, x_2\ldots,x_n\}$ is to be clustered. Each $X \in R^p$ is an attribute vector consisting of p real measurements describing the object. Let objects are to be clustered into non overlapping groups $C = \{C_1, C_2\ldots C_k\}$ (C is known as a clustering), where k is the number of clusters, $C_1 \cup C_2 \cup \ldots \cup C_k = X$, then

$$C_i \neq \phi \text{ for } i=1,\ldots,K,$$

$$C_i \cap C_j = \phi \text{ for } i=1,\ldots,K, j=1,\ldots,K, \text{ and } i \neq j,$$

$$\bigcup_{i=1}^{K} C_i = X .$$

Assign N data points in d dimensions to k clusters, so that a certain underlying notion of homogeneity with in a cluster and heterogeneity across clusters is maximized. The assignment of N points to k clusters is purely a combinatorial one [23], given by

$$NW(N,k) = \frac{1}{k!}\sum_{i=0}^{k}(-1)^i \binom{k}{i}(k-i)^N \qquad (1)$$

Thus, there are a large number of possible partitions even for moderate N and k (e.g. NW (25, 5) ≈ 2.5×10^{15}), and the complete enumeration of every possible partition is simply not possible [2]. This problem is NP-complete and attempting to find a global optimum is usually not computationally viable even moderately sized datasets.

In other words, it is not easy to find the best partitioning even assuming that k is known. Indeed, this is rarely the case in practice. A usual approach is to run a clustering algorithm several times and, based on the obtained results; choose the value for k that provides the most natural clustering.

Although K-Means is one of the widely used clustering techniques. But this algorithm may fail to converge to a local minimum under certain conditions [18]. Moreover, global solutions of large problems cannot be found with a reasonable amount of computation effort [19]. It is because of these factors that several approximate methods are developed to solve the underlying optimization problem. As described in the next section, GA is one such technique that may be efficiently applied for finding optimal clusters by minimizing the extrinsic clustering metric.

III. Genetic Algorithm with Chromosome Reorganization

In this section, a new clustering algorithm based on the proposed chromosome reorganization method and GA is described in details. It includes determination of the number of clusters as the appropriate clustering of the data set. This genetic clustering technique is referred as genetic algorithm with chromosome reorganization (GACR).

A. Chromosome representation

Each of chromosomes is made up of sequence of genes from certain alphabet which is consists of binary digits, floating point numbers, integers and symbols. In proposed clustering algorithm with the string-of-group encoding strategy, each candidate clustering solution is coded as an integer string and the value of an integer in the string represents the label of the group in which the instance is classified [22]. For example, if the data set has six instances $\{x_1, x_2, x_3, x_4, x_5, x_6\}$ the chromosome (1 2 2 2 1 1) represents that the instances $\{x_1, x_5, x_6\}$ are classified in one group, while the instances $\{x_2,x_3,x_4\}$ are classified in the other group, and the partition of the data represented by the chromosome is $\{x_1, x_5, x_6\}$, $\{x_2,x_3,x_4\}$.The use of simple encoding systems causes problem of context insensitivity. The context insensitivity occurs if one clustering solution can be coded by several different chromosomes. For example, following chromosomes represent the same clustering solution:

chromosome 1: 1 1 2 2 3 3
chromosome 2: 2 2 3 3 1 1
chromosome 3: 3 3 2 2 1 1

Where instances $\{x_1, x_2\}$, $\{x_3, x_4\}$ and $\{x_5, x_6\}$ form three clusters. It is irrelevant to the clustering solution if cluster formed by $\{x_1, x_2\}$ is called cluster2 instead of cluster 3 or cluster1.Another problem with this encoding scheme is clustering invalidity. The clustering invalidity occurs if the recombination operator reproduces new clustering solutions, whose number of clusters is smaller than the given number of clusters. For example, if the simple one-point crossover operator is executed on the chromosome (1 1 2 2 3 3) and the chromosome (3 1 1 3 2 2), both new clustering solutions (1 1 2 2 2 2) and (3 1 1 3 3 3) have only two clusters and both of them are invalid. These problems are solved by the developed chromosome reorganization (CR) method. The pseudo code of CR method is given below in Fig.1.

```
Let P1 = {1,2,…,K} and P2 = Φ:
 Set q = 1
  for  r = 1 to N
    if P2(r) = 0 then
      Set s = P1(r)
       for t = 1 to N
          if(P1(t) = s and P2(t) = 0)
             P1(t) = q and P2(t) = 1
          end_if
       end_for
        q = q+1
    end_if
  end_for
```

Fig.1. Pseudo-code of the CR method

The basic steps of genetic algorithm with chromosome reorganization (GACR) are following:

B. Fitness function

The fitness function f of the chromosome is computed as inverse of SSE (sum of squared error):

$$f = \frac{1}{SSE} \qquad (2)$$

The SSE is defined as

$$SSE = \sum_{C_i}\sum_{X\in Ci}(x-m)^T(x-m) = \sum_{Ci}\sum_{x\in Ci}\|x-m_i\|^2 \qquad (3)$$

Begin

1. t = 0
2. Initialize population $P(t)$, /* Popsize = $|P|$ */
3. for i = 1 to Popsize call CR_method for $P(i)$ and evaluate the fitness values of chromosome $P(i)$
4. $t = t + 1$
5. If termination criterion achieved go to step 10
6. Select (P)
7. crossover (P)
8. mutate (P)
9. go to step 3
10. Output best chromosome and stop

End

Fig.2. Basic steps of GACR algorithm

Where x Є C_i is a data point assigned to that cluster. This measure computes the cumulative distance of each pattern from its cluster center of each cluster individually, and then sums those measures over all clusters.

C. Genetic Operators

1) Selection

The selection process selects chromosomes from the mating pool directed by the survival of the fittest concept of natural genetic systems. In proposed GACR, chromosomes selections are based on roulette wheel selection method [22].

2) Crossover

Crossover is a probabilistic process that exchanges information between two parent chromosomes for generating two child chromosomes. The classical crossover operator can not perform well enough due to the problems of clustering invalidity and context insensitivity. It may be necessary to check that offspring produced by a certain operator are valid and reject any invalid chromosomes. In this algorithm a new relation based crossover is presented. The relation based crossover works by building a relation between two parents chromosomes. It works in the following way. First, two chromosomes are selected.

chromosome 1: 22314213
chromosome 2: 31142311

Then, assuming that the clusters 1 and 4 of chromosome1 are randomly selected by GACR.These clusters are copied into corresponding genes of Chromosome2.

chromosome 2: 31114311

The clusters of unchanged genes of chromosome 2 are placed based on relation between selected clusters of chromosome 1 and clusters of corresponding genes of chromosome 2.by applying this procedure, we get new child chromosome 3.

1→4→2

chromosome 3: 32214322

The same procedure is applied to get child chromosome 4, but now considering that the changed clusters of chromosome 2 are copied into chromosome 1.

chromosome 4: 44342213

Crossover probability is selected adaptively as in [30]. The expressions for crossover probabilities are computed as follows. Let f_{max} be the maximum fitness value of the current population, $\overline{f}$ be the average fitness value of the population and f' be the larger of the fitness values of the solutions to be crossed. Then the probability of crossover, p_c, is calculated as

$$p_c = k_1 \times \frac{(f_{\max} - f')}{(f_{\max} - \overline{f})} \quad \text{if } f' > \overline{f}.$$

$$p_c = k_3, \text{ if } f' \le \overline{f}.$$

Here, as in [30], the values of k_1 and k_3 are kept equal to 1.0. Note that, when $f_{\max} = \overline{f}$, then $f' = f_{max}$ and p_c will be equal to k_3. The aim behind this adaptation is to achieve a trade-off between exploration and exploitation in a different manner. The value of p_c is increased when the better of the two chromosomes to be crossed is itself quite poor. In contrast when it is a good solution, p_c is low so as to reduce the likelihood of disrupting a good solution by crossover.

3) Mutation

Each chromosome undergoes mutation with a probability p_m. The mutation probability is also selected adaptively for each chromosome as in [30]. The expression for mutation probability, p_m is given below:

$$p_m = k_2 \times \frac{(f_{\max} - f)}{(f_{\max} - \overline{f})} \quad \text{if } f > \overline{f}.$$

$p_m = k_4$ if $f \le \overline{f}$.

Here, values of k_2 and k_4 are kept equal to 0.5. This adaptive mutation helps GA to come out of local optimum. When GA converges to a local optimum, i.e., when $f_{\max} - \overline{f}$ decreases, p_c and p_m both will be increased. As a result GA will come out of local optimum. It will also happen for the global optimum and may result in disruption of the near-optimal solutions. As a result GA will never converge to the global optimum. But as p_c and p_m will get lower values for high fitness solutions and get higher values for low fitness solutions, while the high fitness solutions aid in the convergence of the GA, the low fitness solutions prevent the GA from getting stuck at a local optimum. The use of elitism will also keep the best solution intact. For a solution with the maximum fitness value, p_c and p_m are both zero. The best solution in a population is transferred undisrupted into the next generation. Together with the selection mechanism, this may lead to an exponential growth of the solution in the population and may cause premature convergence. To overcome the above stated problem, a default mutation rate (of 0.01) is kept for every

solution in the GACR.We have used the mutation operation similar to that used in GA based clustering [30]. In GACR, the best string seen up to the last generation provides the solution to the clustering problem. Elitism has been implemented at each generation by preserving the best string seen up to that generation in a location outside the population. Thus on termination, this location contains the centers of the final clusters.

D. Termination Criterion

We have executed the GACR algorithm for a fixed number of generations. The best string of the last generation provides the solution to the clustering problem. Moreover, the elitist model of GAs has been used, where the best string seen so far is stored in a location within the population.

IV. Experimental Results

The performance of the GACR algorithm, GCA algorithm and K-means algorithms are compared through the experiments based on several artificial data sets and real data sets that are used. All experiments were run on a PC with a 2.0GHz processor and 2 GB RAM. In the experiments, the population size is taken as 40.The crossover and mutation probabilities for GACR-clustering and CGA-clustering algorithm are p_c=0.8 and p_m=0.001, respectively.

A. Artificial data sets

In order to evaluate the performance of the proposed GACR algorithm more objectively, some artificial data sets are generated and performed clustering by using the proposed algorithm.

Data set 1: This is a 5 dimensional data set generated using data generator consists of three classes of 400 data points. This dataset is to be clustered into 4 clusters. The final clustering results obtained by K-mean, GCA and GACR are given in figures 3(a), (b) and (c), respectively.

Data set 2: This is a 10-dimensional data set consists of eight clusters of 1000 data points. This dataset is to be clustered into 8 clusters. The final results corresponding to K-mean, GCA and GACR are shown in figures 4(a), (b) and (c), respectively.

Data set 3: This is a 50 dimensional data set consists of 300 data points. This dataset is to be clustered into 2 clusters. Figures 5(a), (b) and (c) show the results for K-means, GCA and GACR, respectively.

It is to be noted that for each of the above three data sets, The K-means is unable to provide the correct clustering. However, the GCA is correctly clustered the data points of data set 1, but it is unable to detect the correct clusters in data set 2 and dataset 3.The GACR is able to detect the clusters reasonably well in all data sets.

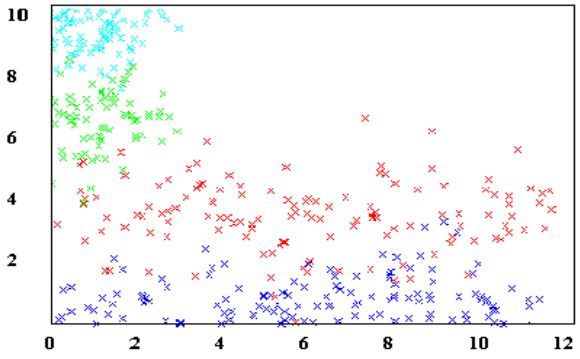

Fig. 3(a) Clustering of Data1 obtained by K-means

B. Real Data sets

The performance of the GACR algorithm, CGA algorithm and K-means algorithm are compared through the experiments on the following three real data sets.

Iris: Iris data set consists of 150 data points distributed over three clusters. Each cluster has 50 points. This data set represents different categories of irises characterized by four feature values. It has three classes Setosa, Versicolor and Virginica among which the last two classes have a large amount of overlap while the first class is linearly separable.

Cancer: Wisconsin Breast cancer data set consisting of 683 sample points. Each pattern has nine features corresponding to clump thickness, cell size uniformity, cell shape uniformity, marginal adhesion, single epithelial cell size, bare nuclei, bland chromatin, normal nucleoli and mitoses. There are two categories in the data: malignant and benign. The two classes are known to be linearly separable.

***Wine*:** This is the Wine recognition data consisting of 178 instances having 13 features resulting from a chemical analysis of wines grown in the same region in Italy but derived from three different cultivars. The analysis determined the quantities of 13 constituents found in each of the three types of wines.

The performance results reported in Table 1, clearly demonstrate the clustering accuracy of K-means, GCA and GACR for artificial and real data sets. Table 2 indicates the computing times of K-means, GCA and GACR for clustering of above data sets. The GACR perform better results in terms of reduction in CPU time in compared to GCA and K-means.

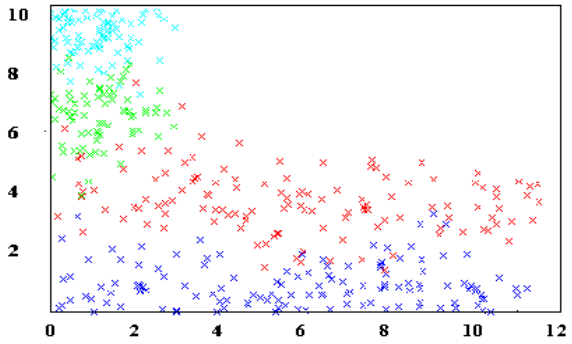

Fig. 3(b) Clustering of Data1 obtained by GCA

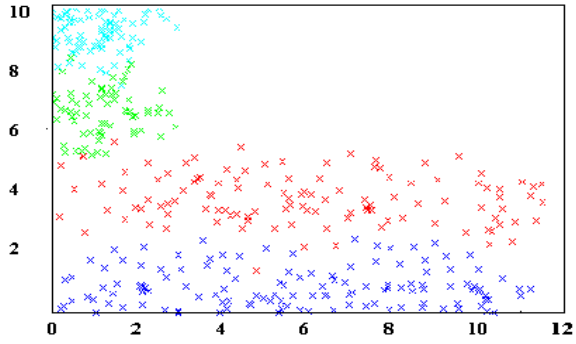

Fig. 3(c) Clustering of Data1 obtained by GACR

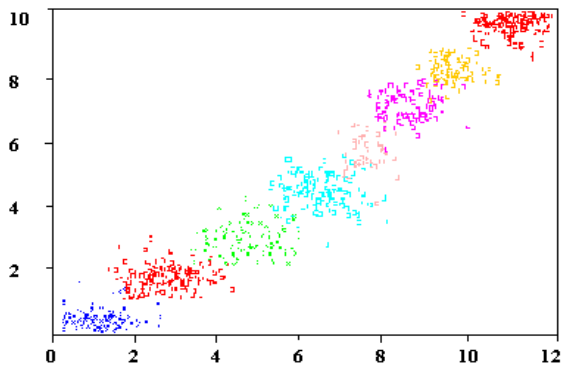

Fig. 4(a) Clustering of Data2 obtained by K-means

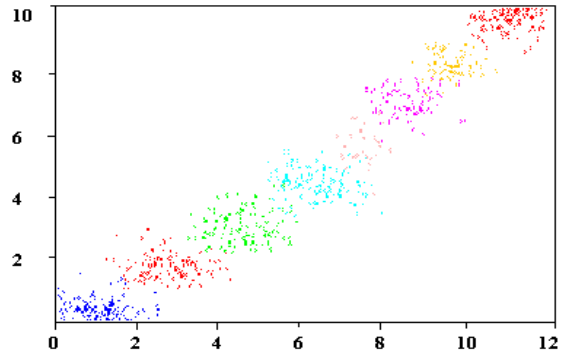

Fig. 4(b) Clustering of Data2 obtained by GCA

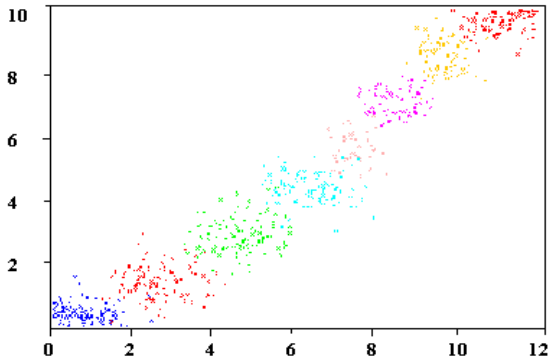

Fig. 4(c) Clustering of Data2 obtained by GACR

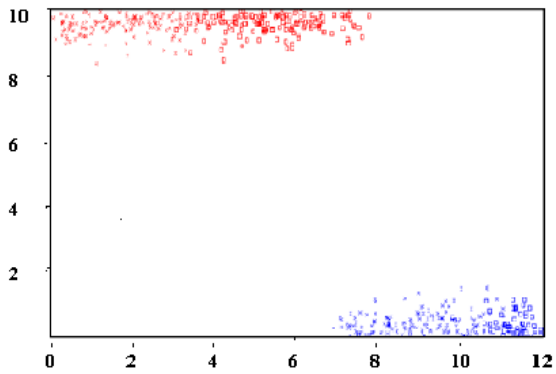

Fig. 5(a) Clustering of Data3 obtained by K-means

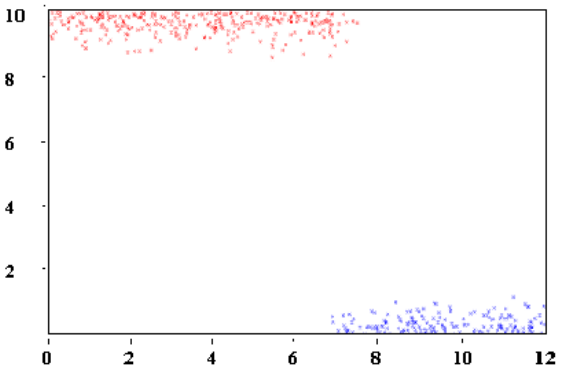

Fig. 5(b) Clustering of Data3 obtained by GCA

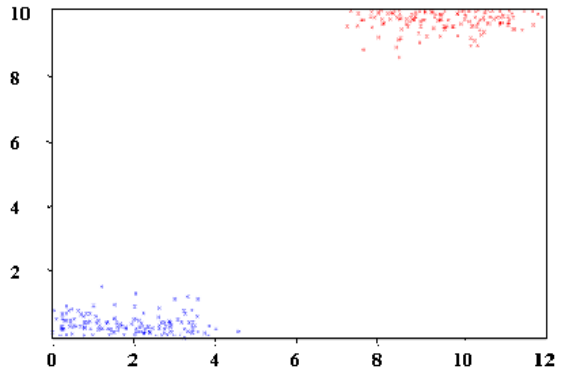

Fig. 5(c) Clustering of Data3 obtained by GACR

TABLE 1. ACCURACY OF K-MEANS, GCA AND GACR

Data sets	Accuracy of clustering in percentage		
	K-Mean	GCA	GACR
Data set 1	78.2	95.7	96.6
Data set 2	71.0	82.1	96.5
Data set 3	60.8	75.9	85.4
Iris	87.4	97.0	99.0
Cancer	85.5	94.2	96.8
Wine	86.0	92.5	95.8

TABLE 2. COMPUTING TIMES OF K-MEANS, GCA AND GACR

Data sets	Computing times in seconds		
	K-Mean	GCA	GACR
Data set 1	84	20	7
Data set 2	196	48	15
Data set 3	152	38	10
Iris	32	7	2
Cancer	105	25	6
Wine	68	16	4

V. Conclusion

In this paper a genetic clustering technique (GACR) is proposed. In order to reduce context insensitivity caused by different chromosomes describing the same cluster result, a chromosome reorganization method has been presented. A new relation based crossover operator has also defined. These two new approaches allow our GACR to explore the search space more effectively. The GACR used adaptive probabilities of crossover and mutation to prevent the GACR clustering algorithm from getting stuck at a local optimal solution.

Experimental results on different data sets demonstrate the comparison of GACR algorithm, GCA algorithm and K-means algorithm. GACR is found to provide satisfactory performance where GCA and K-means fails. The superiority of GACR is also established on three real-life data sets. These real-life data sets are of different characteristics, with the number of dimensions varying from 4 to 20. Results on various artificial and real-life data sets demonstrate that GACR has better performance than GCA and K-means clustering algorithm.

Acknowledgment

The authors gratefully acknowledge Prof. P K Das Dean, Faculty of Engineering and Technology, MITS, Rajasthan, India for his kind support in this research work.

Reference

[1] Jiawei Han and M.kamber.Data mining: Concepts and Techniques,Morgan Kaufmann,2004.

[2] B. Everitt,S.Landau,M.Leese, Cluster Analysis, Arnold, London, 2001 .

[3] R. Xu, D. Wunsch,"Survey of clustering algorithms," IEEE Trans.Neural Networks ,vol. 16 ,2005,pp. 645–678.

[4] J.Handl, J.Knowles, D. B. Kell, "Computational cluster validation in post- genomic data analysis," Bioinformatics, vol. 21, 2005, pp. 3201–3212.

[5] M.Ester,H.P.Kriegel,J.Sander, "A density-based algorithm for discovering clusters in large spatial databases with noise,"in Proceedings of the 2nd International Conference on Knowledge Discovery and Data Mining,1996, pp. 226–231.

[6] S.Y.Lu,K.S.Fu,"A sentence-to-sentence clustering procedure for pattern analysis," IEEE Trans.Syst.ManCybern.,vol. 8,1978,pp.381–389.

[7] G.J.Mclachlan,K.E.Basford,Mixture Models: Inference and Applications to Clustering, Marcel Dekker, NewYork, 1988.

[8] G.J.Mclachlan,T.Krishnan,The EM Algorithm and Extensions,Wiley,New York, 1997.

[9] Singh Vijendra, Sahoo Laxman and Kelkar Ashwini,"Mining Subspace Clusters in High Dimensional Data," International Journal of Recent Trends in Engineering,vol. 3,2010,pp.118-122.

[10] Singh Vijendra, Sahoo Laxman and Kelkar Ashwini,"An Effective Clustering Algorithm for Data Mining," in proceedings of IEEE International conference on Data Storage and Data Engineering,2010,pp.250-253.

[11] S.Mitra,H.Banka,"Multi-objective evolutionary biclustering of gene expression data," Pattern Recognition vol.39 ,2006,pp. 2464–2477.

[12] S.C.Madeira,A.L.Oliveira,"Bi-clustering algorithms for biological data analysis:a survey, "IEEETrans. Comput. Biol. Bioinf. ,vol. 1, 2004,pp. 24–45.

[13] Y.Kluger,R.Barsi,J.T.Cheng,M.Gerstein, "Spectral biclustering of micro array data:co clustering genes and conditions ,"Genome Res.,vol.13,2003,pp. 703–716.

[14] Y.Hong,S.Kwong,Y.C.Chang,Q.S.Ren,"Unsupervised feature selection using clustering ensembles and population based incremental learning algorithm,"Pattern Recognition ,vol. 41,2008,pp. 2742–2756.

[15] A.Strehl,J.Ghosh,"Cluster ensembles—aknowledge reuse frame work for combining ultiple partitions, "J.Mach. Learn. Res.,vol. 3,2002,pp. 583–617.

[16] A.Topchy,A.K.Jain,W.Punch,"Clustering ensembles models of consensus and k partitions," IEEE Trans .Pattern Anal. Mach. Intell. Vol. 27,2005.

[17] M.B.Dale,P.T.Dale,"Classification with multiple dissimilarity matrices," Coenoses vol.9,1992,pp.1–13.

[18] S.Z.Selim, M.A.Ismail, "K-means type algorithms: a generalized convergence theorem and characterization of local optimality," IEEE Trans. Pattern Anal. Mach. Intell., vol.6,(1984,pp.81–87.

[19] H.Spath, Cluster Analysis Algorithms, Ellis Horwood, Chichester, UK, 1989.

[20] T.Kanungo, D.Mount, N.S. Netanyahu, C.Piatko, R.Silverman, A.Wu, "An efficient K-means clustering algorithm: analysis and implementation," IEEE Trans. Pattern Anal. Mach. Intell. , vol.24, 2002, pp.881–892.

[21] A.Likas, N.Vlassis, J.J.Verbeek, "The global K-means clustering algorithm, "Pattern Recognition vol.36, 2003, pp.452–461.

[22] Goldberg, D.E., Genetic Algorithms in Search, Optimization and Machine Learning. Addison-Wesley, 1989.

[23] G.L. Liu, Introduction to Combinatorial Mathematics, Mc Graw Hill, New York, 1968.

[24] C.A.Murthy and N.Chowdhury,"In search of optimal clusters using genetic algorithms,"pattern Recog.lett.,vol.17, 1996,pp.825-832.

[25] L. Y. Tseng and S. B. Yang,"A genetic approach to the automatic clustering algorithm, Pattern Recognition," vol. 34, no. 2,2001, pp. 415-424.

[26] S. Bandyopadhyay and U. Maulik.,"An evolutionary technique based on K-means algorithm for optimal clustering in RN," Information Sciences,vol. 146, no.1-4, 2002,pp. 221-237.

[27] Lai.,"A novel clustering approach using hierarchical genetic algorithms, "Intelligent Automation and Soft Computing, vol. 11, no. 3, 2005,pp. 143-153.

[28] H. J. Lin, F. W. Yang and Y. T. Kao.An efficient GA-based clustering technique, Tamkang Journal of Science and Engineering, vol. 8, no. 2,pp. 113-122, 2005.

[29] Sanghamitra Bandyopadhyay, Sriparna Saha, "A clustering method using a new point symmetry-based distance measure," Pattern reconition ,vol. 40 ,2007,pp. 3430 – 3451.

[30] M.Srinivas, L.Patnaik," Adaptive probabilities of crossover and mutation in genetic algorithms, "IEEE Trans. Syst. ManCybern. Vol. 24,1994,pp. 656–667.

Vijendra Singh received the M.Tech degree in Computer Science and Engineering from Birla Institute of Technology, Mesra (Ranchi), INDIA. His research interests include Data Mining, Pattern Recognition, Evolutionary and Soft Computation and Bioinformatics.

Laxman Sahoo received the M.Tech degree in Computer Science and Engineering and the PhD degree awarded from Indian Institute of Technology, Khargpur (INDIA).His field of interest includes data mining, rough sets and Pattern Recognition.

Ashiwin Kelker received M.Sc (Computer Science) from Indian Institute of Technology, Mumbai (INDIA), and MS and awarded PhD degree from Arizona State University, Arizona (USA).Her research interests include Data Mining and Rough Sets.