Data Clustering with Mixed Features by Multi Objective Genetic Algorithm

Dipankar Dutta
Department of
Computer Science and
Information Technology,
University Institute of Technology,
The University of Burdwan,
Golapbug (North), Burdwan,
West Bengal, India
PIN-713104
Email:dipankar_dutta@rediffmail.com

Paramartha Dutta
Department of Computer
and System Sciences,
Visva-Bharati University,
Santiniketan, West Bengal, India
PIN-731235
Email:paramartha.dutta@gmail.com

Jaya Sil
Department of
Computer Science and Technology,
Bengal Engineering
and Science University,
Shibpur, West Bengal, India
PIN-711103
Email:js@cs.becs.ac.in

***Abstract*—In the paper, real coded multi objective genetic algorithm ($MOGA$) based K-clustering method has been studied where K represents the number of clusters known a priori. Proposed method has the capability to deal with continuous and categorical features (mixed features) of data set. Commonly means and modes of features represents clusters for continuous and categorical features respectively. For this reason, K-means and K-modes are most popular clustering algorithm for continuous and categorical features respectively. The searching power of Genetic Algorithm (GA) is exploited to search for suitable clusters and cluster centroids (means or modes) so that intra-cluster distance (Homogeneity, H) and inter-cluster distances (Separation, S) are simultaneously optimized. It is achieved by measuring H and S using a special distance per feature metric, suitable for continuous and categorical features both. We have selected four benchmark data sets from UCI Machine Learning Repository containing continuous and categorical features both. Here, K-means and K-modes is hybridized with GA to combine global searching capabilities of GA with local searching capabilities of K-means and K-modes. Considering context sensitivity, we have used a special crossover operator called "pairwise crossover" and "substitution".**

I. Introduction

Data mining is an interdisciplinary field of research applied to extract high level knowledge from real-world data sets [15]. Clustering happens to be one of the important data mining task.

Clustering is an unsupervised learning process where class labels are not available at the training phase.

GA is "search algorithms based on the dynamics of natural selection and natural genetics" [19]. Categories of GAs are - simple GA (SGA) and multi objective GA ($MOGA$). When an optimization problem involves only one objective, the task of finding the best solution is a single objective optimization problem. However, in the real world, life appears to be quite complex. Most of the problems are consisting of more than one interdependent objective, which are to be minimized or maximized simultaneously. These types of problems are multi objective optimization problem [6]. The use of GAs in clustering is an attempt to exploit effectively the large search space usually associated with cluster analysis and better interactions among the features to form chromosomes.

A cluster is a set of entities that are alike, and entities from different clusters are not alike [13]. From optimization viewpoint and due to its unsupervised nature, clustering is one of the most challenging problems, considered as a NP-hard grouping problem [14]. Evolutionary algorithms like GAs are efficient to solve the NP-hard problems with high degree of complexity to find out near-optimal solutions to such problems in reasonable time. Although GAs have been used in data clustering problems, most of them are single objective in nature. Naturally, this is hardly equally applicable to all kinds of data sets. So to solve many real world problems like clustering, it is necessary to optimize more than one objective simultaneously by $MOGA$. As the relative importance of different clustering criteria are unknown, it is better to optimize Homogeneity (H) and Separation (S) separately rather than combining them into a single measure to be optimized. Objects within the same clusters are similar implying low intra-cluster distances (H) and at the same time, objects belonging to different clusters are dissimilar, thereby inducing high inter-cluster distances (S). In the paper, clustering is considered as an intrinsically multi objective optimization problem [24] by involving homogeneity and separation together. However, choosing optimization criterion is one of the fundamental dilemmas in clustering [36].

This paper is organized as follows. Section II reviews previous relevant works. Section III clarifies definitions pertaining to the problem. Section IV presents the concept and the structure of the proposed algorithm. Section V, contains experimental results with real life benchmark data sets. Finally, section VI we present conclusions and some remarks on the future research.

978-1-4673-5115-7

II. Previous Work

Clustering is one of the most studied areas of data mining research. Extensive surveys of the clustering problems and algorithms are presented in [16], [21], [27], [32], [46]. With the growing popularity of soft computing methods, researchers of late extensively use intelligent approaches in cluster analysis of data sets.

Clustering techniques can be generally divided into two categories: hierarchical and non-hierarchical [13], [32]. Partitioning methods falls into the category of non-hierarchical. In this method data sets are divided among K many partitions or clusters where $K \leq m$, and m is the number of objects or tuples in the data sets. In this category, most popular methods are K-means [47], [48], fuzzy C-means (FCM) [3], K-modes [28], [29], fuzzy K-modes [30] and K-medoids [35]. Among these K-means and FCM is applied in continuous domains. K-modes and fuzzy K-modes are popular methods in categorical domain. K-prototype [28] is well known for mixed type of data. However, final solutions given by these algorithms depends upon initial solution and these may converge to a local optimal solution.

GA has been used for clustering to find a globally optimal clustering solution. Initially GAs were used to deal with single objective optimization problems such as minimizing or maximizing a variable of a function. Such GAs is simple GA (SGA). David Schaffer proposed vector evaluated genetic algorithm [49], which was the first inspiration towards $MOGA$. Several improvements on the vector evaluated genetic algorithm are proposed in [19]. Vector evaluated genetic algorithm dealt with multivariate single objective optimization. V. Pareto suggested Pareto approach [45] to cope with multi-objective optimization problems. In this work, chromosomes lying on the Pareto optimal front (dominant chromosomes) are selected by $MOGA$, providing near optimal solutions for clustering. According to the dominance relation between two candidate chromosomes, a candidate clustering chromosome Ch_1 dominates another candidate clustering chromosome Ch_2 if and only if: 1) Ch_1 is strictly better than Ch_2 in at least one of all the measures considered in the fitness function and 2) Ch_1 is not worse than Ch_2 in any of the measures considered in the fitness function. In this paper, we have developed our own $MOGA$ to find out near optimal clusters by optimizing H and S values. From the viewpoint of the chromosome encoding scheme $MOGA$ can be categorized in binary coded $MOGA$ and real coded $MOGA$. In order to keep the mapping between the actual cluster centroids and the encoded centroids effective, real coded $MOGA$ has been implemented in the work.

Most of the clustering approaches, whether single objective or multi-objective, are converted to a single objective by weighted objective functions [34] or by any other means. However, only a few $MOGA$ clustering approaches have been proposed so far and their experimental results have demonstrated that $MOGA$ clustering approaches significantly outperform existing single objective GA clustering approaches [22], [23]. Earlier work on clustering by the single objective genetic algorithm are [9], [17], [20], [39], [41], [44], [50] and by $MOGA$ are [1], [8], [11], [12], [18], [25], [37], [38], [40], [42], [43]. Among them, [1], [9], [11], [39] are works on continuous features and [8], [12], [18], [25], [40], [42], [43] are on categorical features. Very few works [4], [33], [28], [31], [51] has been reported so far which can deal with continuous and categorical features both. But in [4], [28], [31] researchers have not done global optimization. In [33], [51] have done global optimization by SGA and not by $MOGA$. Hence, finding globally optimal clustering solutions involving more that one objectives (two objectives) by $MOGA$ for mixed attributes are main contribution of this work. Davies-Bouldin (DB) Index [5] has been used to compare performance of clustering algorithms. It shows the superiority of global optimization capability of proposed method over local search algorithm like K-means or K-modes. Evaluating the goodness of clusters by other measures (DB Index) rather than using measures used for optimization by $MOGA$ (H and S) is common in literature [1], [11], [12], [25], [40], [42], [43]. If we use DB Index as optimization measure in place of H and S, then it is a conversion of multi objective optimization problem into single objective optimization problem. Finding a solution by $MOGA$ can be treated as a two stepped process. In the first step $MOGA$ will find out a set of near-optimal solutions close to Pareto-optimal front. In second step a solution is selected from the set of solutions using higher order knowledge [6]. We are following this approach in this work.

III. Definitions

A relation R and a list of features $A_1, A_2, ...A_n$ defines a relational schema $R(A_1, A_2, ...A_n)$, where n = total number of features.

The domain of A_i is $dom(A_i)$, where $1 \leq i \leq n$. A_i^{con} denotes attributes with continuous domain and A_i^{cat} denotes attributes with categorical domain.

X represents a data set comprising a set of tuples. That is $X = \{t_1, t_2, ...t_m\}$, where m = total number of tuples or records.

Each tuple t is an n-dimensional attribute vector, i.e., an ordered list of n values i.e. $t = [v_1^t, v_2^t, ...v_n^t]$, where $v_i^t \in dom(A_i)$, with $1 \leq i \leq n$. v_i^t is i^{th} value in tuple t, which corresponds to feature A_i. Each tuple t belongs to a predefined class, represented by v_n^t where $v_n^t \in dom(A_n)$. Data treated in clustering are usually unlabeled. So, $t = [v_1^t, v_2^t, ...v_{n-1}^t]$.

Formally, the problem is stated as every t_i of X ($1 \leq i \leq m$) is to be clustered into K number of non-overlapping groups $\{C_1, C_2, ...C_K\}$; where $C_1 \cup C_2 \cup ...C_K = X$, $C_i \neq \emptyset$, and $C_i \cap C_j = \emptyset$ for $i \neq j$ and $j \leq K$.

The solution of the clustering problem may be a set of centroids (CE) that is $\{CE_1, CE_2, ...CE_K\}$. $(n-1)$-dimensional feature vector, that is $[c_1^i, c_2^i, ...c_{n-1}^i]$ represents CE_i.

For all A_i^{con}, A_i^{min} (minimum value of A_i) and A_i^{max} (maximum value of A_i) are calculated, where $1 \leq i \leq n-1$ to convert A_i^{con} in normalized value by using the following simple linear interpolation based conversion formula.

$$nv_i^t = (v_i^t - A_i^{min})/(A_i^{max} - A_i^{min}) \text{ for } i = 1, 2, ...(n-1). \quad (1)$$

Equations (2), (3) and (4) calculate distance per feature between one cluster centroid and one tuple, two cluster centroids and two tuples respectively.

$$d(CE_i, t_j) = [\sum_{l=1}^{n-1}[(c_l^i - nv_l^j)^2 + MOD(c_l^i, v_l^j)]/(n-1)]^{1/2} \quad (2)$$

$$d(CE_i, CE_j) = [\sum_{l=1}^{n-1}[(c_l^i - c_l^j)^2 + MOD(c_l^i, c_l^j)]/(n-1)]^{1/2} \quad (3)$$

$$d(t_i, i_j) = [\sum_{l=1}^{n-1}[(nv_l^i - nv_l^j)^2 + MOD(v_l^i, v_l^j)]/(n-1)]^{1/2} \quad (4)$$

For continuous feature we are using Euclidean distance by using normalized values (nv_i^t) and for categorical feature we are using Mod distance (MOD). $MOD(c_l^i, v_l^j)$, $MOD(c_l^i, c_l^j)$ and $MOD(v_l^i, v_l^j)$ are all equal to 0 if $c_l^i = v_l^j$, $c_l^i = c_l^j$ and $v_l^i = v_l^j$. Otherwise they are equal to 1.

IV. Proposed Approaches

Original data sets are rearranged to label the last feature as class label and so class label becomes the n^{th} feature. In clustering, class labels are unknown so we are considering first $(n-1)$ features of data sets. In case of Acute Inflammations data set [26] last two features are class labels. So in this case we are considering first $(n-2)$ features of data set.

1) Building initial population: In most of the literature on GA, fixed number of prospective solutions build initial population (IP). Here, the size of IP is the nearest integer value of 10% of total number of tuples (m) in the data set. Although correlation between IP size and the number of instances in the data set are not explicitly known, our experience suggests that IP size guides searching power of GA and therefore its size should increase with the size of the data set. K number of tuples are chosen randomly to constitute j^{th} chromosome Ch_j where $1 \leq j \leq$ initial population size (IPS) in a population. The process is repeated for IPS number of times. Each chromosome represents a prospective solution, which is a set of cluster centroids (CEs) of K clusters.

As tuples of data set build cluster centroids, it induces faster convergence of $MOGA$, compared to building chromosomes by randomly choosing feature values from the same feature domains.

2) Reassign cluster centroids: Using every chromosome, a set of CEs, represented as, $\{CE_1, CE_2, ...CE_K\}$ are randomly initialized. A combination of K-means and K-modes algorithm produces a set of new cluster centroids that is $\{CE_1^*, CE_2^*, ...CE_K^*\}$, which forms new chromosomes. Only one iteration of combination of K-means and K-modes algorithm is running. K-means is applicable for A_i^{con} and K-modes is applicable for A_i^{cat} for obvious reason.

3) Crossover: Chromosomes generated by combination of K-means and K-modes algorithm are input to the crossover. Context insensitivity is an important issue in a grouping task like clustering. Meaning of context insensitivity is "the schemata defined on the genes of the simple chromosome do not convey useful information that could be exploited by the implicit sampling process carried out by a clustering GA" [14]. In their survey paper Hruschka et. al. [27] shows drawback of conventional single point crossover operators considering context sensitivity. We have used a special crossover operator called "Pairwise crossover" described by Fränti in [17] as "The clusters between the two solutions can be paired by searching the "nearest" cluster (in the solution B) for every cluster in the solution A. Crossover is then performed by taking one cluster centroid (by random choice) from each pair of clusters. In this way we try to avoid selecting similar cluster centroids from both parent solutions. The pairing is done in a greedy manner by taking for each cluster in A the nearest available cluster in B. A cluster that has been paired cannot be chosen again, thus the last cluster in A is paired with the only one left in B." This algorithm does not give the optimal pairing (2-assignment) but it is a reasonably good heuristic for the crossover purpose. Crossover probability (P_{co}), chosen as 0.9 that is 90% chromosomes undergo with this crossover.

4) Substitution: Substitution probability (P_{sb}) of $MOGA$ is 0.1. Child chromosomes are produced by crossover are parent chromosomes of this genetic operator. In conventional mutation, any random value from $dom(A_i)$ can replace v_i, resulting different chromosomes. Considering context sensitivity, instead of replacement of any v_i of chromosome, substitution is replacing cluster centroids by any tuples randomly. Approximately number of substitution (N_{sb}) = $\lfloor (N_{pch} \times K \times P_{sb}) \rfloor$.

5) Combination: At the end of every generation of $MOGA$ some chromosomes are lying on the Pareto optimal front. These chromosomes have survived and these are known as Pareto chromosomes. Chromosomes obtained from previous substitution method, say NC_{pre} and previous generation Pareto chromosomes of $MOGA$, say PC_{pre} are considered to perform in the next generation. For the first generation of GA, PC_{pre} is zero because of the nonexistence of previous generation.

In general, for i^{th} generation of $MOGA$, if NC_{pre} is m and PC_{pre} at the end of $(i-1)^{th}$ generation of $MOGA$ is n then after combination, number of chromosomes are $n+m$ or $PC_{pre} + NC_{pre}$.

6) Eliminating duplication: Survived chromosomes from previous generation may be generated again in the present generation of $MOGA$, resulting multiple copies of the same chromosome. So in this step system is selecting unique chromosomes from the combined chromosome set. One may argue with the requirement of this step. But if we remove this step then population size will be very high at the end of some

iterations of GA.

7) Fitness Computation: As stated earlier, intra-cluster distance (Homogeneity) (H) and inter-cluster distances (Separation) (S) are two measures for optimization. Maximization of $1/H$ and S are the twin objectives of $MOGA$. It converts the optimization problem befitting into max-max framework. The values of H and S are computed by using distance per feature measures.

Equation 2 is calculating distance per feature between one cluster centroid and one tuple.

If $d_{lowest}(CE_i, t_j)$ is the lowest distance for any value of i (where $1 \leq i \leq K$), then t_j is assigned to C_i. Ties are resolved arbitrarily. Equation 5 defines H_i (average intra-cluster distance or homogeneity of i^{th} cluster).

$$H_i = [\sum_{j=1}^{m_i} d_{lowest}(CE_i, t_j)]/m_i \tag{5}$$

where m_i is the number of tuples belonging to i^{th} cluster and t_j is assigned to i^{th} cluster based on lowest distance.

Summation of H_i is H, defined in equation 6 as

$$H = \sum_{i=1}^{K} H_i \tag{6}$$

Say, t_j and t_p are two distinct tuples from a data set where t_j is assigned to C_x and t_p is assigned to C_y using $d_{lowest}(CE_i, t_j)$. S is defined in equation 7 as

$$S = \sum_{j,p=1}^{m} d(t_j, t_p) \tag{7}$$

where $j \neq p$ and $x \neq y$.

8) Selection of Pareto chromosomes: All chromosomes lying on the front of $max(1/H, S)$ are selected as Pareto chromosomes. The process is not restricting the number of selected chromosomes. In other words, maximizing $1/H$ and S are two objectives of $MOGA$. As this process is applied after combination, elitism [6], [7] is adhered to.

9) Building intermediate population: From 2^{nd} generation onwards, for every even generation, selected Pareto chromosomes of previous generation build the population. This helps to build a population of chromosomes close to the Pareto optimal front resulting fast convergence of $MOGA$. For every odd generation of $MOGA$ or if for any generation previous generation of $MOGA$ does not produce Pareto chromosomes greater than 2 then population are generated using procedure of generating initial population discussed earlier in section IV-1. This introduces new chromosomes in population and induces diversity in the population. Population size also varies from generation to generation.

From the above discussion, one may get wrong idea that we are destroying any search we are carrying out during a generation and there is no evolution of solution. In other words, there is really no time for evolution. But notice that, we are preserving Pareto chromosomes at every generation (i^{th} generation of GA) and that is combined with present generation chromosomes ($(i+1)^{th}$ generation of GA) during Combination described in section IV-5. Selection of Pareto chromosomes are done on combined population set. As pointed out, that at every odd generation we generate the population applying the same procedure through which the initial population was generated. But is these cases we are not loosing Pareto chromosomes selected by previous generation. We stores them separately to combine them during Combination (described in section IV-5).

V. Testing and Comparison

In literature different validity indices are available to measure the quality of clusters. Davies-Bouldin (DB) Index [5] is used here. Equation 8 is calculating DB Index.

$$DB\ Index = 1/K \sum_{i=1, i\neq j}^{K} max((H_i + H_j)/d(CE_i, CE_j)) \tag{8}$$

where K is the number of clusters, H_i is the average distance of all patterns in cluster i to their cluster centroid CE_i, H_j is the average distance of all patterns in cluster j to their cluster mode CE_j and $d(CE_i, CE_j)$ is the distance of cluster centroids CE_i and CE_j.

The performance of two clustering algorithms are compared on the basis of the results obtained out of the four popular data sets from UCI Machine Learning Repository [26]. Table I summarizes features of data sets.

Table II compares results obtained by the two clustering algorithms. For every data set and for every clustering algorithm, average values of 10 individual runs are tabulated. As combination of K-means and K-modes algorithm is dealing with only one solution A3 is equivalent to B3, B4, B5 of $MOGA(H, S)$.

Following important observations are summarized here, based on the analysis of table II. H for all data sets in column B1 is lower than the values in column A1 because combination of K-means and K-modes is doing local optimization or it locally optimizes intra cluster distance and $MOGA(H, S)$ globally optimizes intra cluster distance. Result shows the superiority of GA over combination of K-means and K-modes. Lowest DB Index of the population (column B6) of the proposed $MOGA(H, S)$ is much better than DB Index obtained by combination of K-means and K-modes (column A3).

VI. Conclusions

In this work, we have implemented a novel real coded hybrid elitist MOGA for K-clustering ($MOGA(H, S)$). It is known that elitist model of GAs provides the optimal string as the number of iterations increases to infinity [2], [10].

We have achieved one important data mining task of clustering by finding cluster centroids. Continuous and categorical features are considered as equally important in this work. Separate optimization objectives of $MOGA$ can be designed for continuous and categorical features. Dealing with missing features values and unseen data are other problem areas. It

TABLE I
FEATURES OF DATA SETS USED IN THIS ARTICLE (HERE, # INDICATES NUMBER OF).

Name	#Tuples	#Continuous Attributes	#Categorical Attributes	#Clusters
Acute Inflammations	120	1	7	2
Liver Disorders	345	6	1	2
Statlog(Heart)	270	7	7	2
Teaching Assistant Evaluation	151	1	5	3

TABLE II
COMPARISON OF PERFORMANCE OF CLUSTERING ALGORITHMS

Meaning of notations used in this table

A1:H of combination of K-means and K-modes, A2:S of combination of K-means and K-modes, A3:DB Index of combination of K-means and K-modes

B1:Optimized min. H of $MOGA(H, S)$ population, B2:Optimized max. S of $MOGA(H, S)$ population, B3:DB Index of chromosome giving optimized min. H of $MOGA(H, S)$ population, B4:DB Index of chromosome giving optimized max. S of $MOGA(H, S)$ population, B5:Average DB Index of $MOGA(H, S)$ population, B6:Best DB Index of $MOGA(H, S)$ population

Methods	Combination of K-means and K-modes			$MOGA(H, S)$					
Data set	A1	A2	A3	B1	B2	B3	B4	B5	B6
Acute Inflammations	1.04	2951.36	1.22	0.82	3258.84	1.23	2.33	1.14	0.95
Liver Disorders	0.26	4663.57	1.38	0.15	6043.81	1.32	1.80	2.23	0.78
Statlog(Heart)	1.07	14469.81	1.27	1.06	14808.36	1.27	1.37	1.263	1.16
Teaching Assistant Evaluation	2.40	6877.20	1.69	1.78	7222.12	1.47	3.36	1.65	1.35

may be interesting to adapt $MOGA$ based K-clustering with unknown number of clusters. Deciding optimum value of K is another research issue. Authors are working in these directions.

REFERENCES

[1] S. Bandyopadhyay, U. Maulik, and A. Mukhopadhyay, "Multiobjective genetic clustering for pixel classification in remote sensing imagery," *IEEE Trans. Geosci. Remote Sens.*, vol. 45, no. 5, May. 2007, pp. 1506-1511, doi:10.1109/TGRS.2007.892604.

[2] D. Bhandari, C. A. Murthy, and S. K. Pal, "Genetic algorithm with elitist model and its convergence," *Int. J. Patt. Recogn. Artif. Intell.*, vol. 10, no. 06, Sep. 1996, pp. 731–747, doi:10.1142/S0218001496000438.

[3] J. C. Bezdek, *Pattern Recognition with Fuzzy Objective Function Algorithms*. New York, NY: Plenum, 1981.

[4] T. Chiu, D. Fang, J. Chen, Y. Wang, and C. Jeris, "A robust and scalable clustering algorithm for mixed type attributes in large database environment," in *Proc. of the 7th ACM SIGKDD international conference on Knowledge discovery and data mining (KDD 2001)*, ACM Press, New York, New York, USA, Aug. 2001, pp. 263–268, doi: 10.1145/502512.502549.

[5] D. L. Davies, and D. W. Bouldin, "A cluster separation measure," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. PAMI-1, no. 2, Apr. 1979, pp. 224-227, doi:10.1109/TPAMI.1979.4766909.

[6] K. Deb, *Multi-Objective Optimization Using Evolutionary Algorithms*. New York: John Wiley & Sons, 2001.

[7] K. Deb, A. Pratap, S. Agarwal, and T. Meyarivan, "A fast and elitist multiobjective genetic algorithm: NSGA-II," *IEEE Trans. Evol. Comput.*, vol. 6, no. 2, Apr. 2002, pp. 182-197, doi:10.1109/4235.996017.

[8] S. Deng, Z. He, and X. Xu, "G-ANMI: A mutual information based genetic clustering algorithm for categorical data," *Knowledge-Based Systems*, vol. 23, no. 2, Mar. 2010, pp. 144-149, doi:10.1016/j.knosys.2009.11.001.

[9] K. Dhiraj, and S. K. Rath, "Comparison of SGA and RGA based clustering algorithm for pattern recognition," *Int. J. of Recent Trends in Engg.*, vol. 1, no. 1, May. 2009, pp. 269–273.

[10] P. Dutta, and P. DuttaMajumder, "Convergence of an evolutionary algorithm," in *Proc. of the 4th Int. Conf. on Soft Computing (IIZUKA 1996)*, World Scientific, Fukuoka, Japan, Sep.–Oct. 1996, pp. 515–518.

[11] D. Dutta, P. Dutta, and J. Sil, "Clustering by multi objective genetic algorithm," in *Proc. of 1st Int. Conf. on Recent Advances in Information Technology (RAIT 2012)*, IEEE Press, Dhanbad, India, Mar. 2012, pp. 548–553, doi:10.1109/RAIT.2012.6194619.

[12] D. Dutta, P. Dutta, and J. Sil, "Clustering Data Set with Categorical Feature Using Multi Objective Genetic Algorithm," in *Proc. Int'l Conf. on Data Science and Engineering (ICDSE 2012)*, IEEE Press, Cochin, Kerala, India, Jul. 2012, pp. 103–108, doi: 10.1109/ICDSE.2012.6281897.

[13] B. S. Everitt, *Cluster Analysis*. London, U.K.: Edward Arnold, 1993.

[14] E. Falkenauer, *Genetic Algorithms and Grouping Problems*. New York, NY: John Wiley & Sons, 1998.

[15] A. A. Freitas, "A Survey of Evolutionary Algorithms for Data Mining and Knowledge Discovery," in *Advances in Evolutionary Computing of Natural Computing Series*, Chapter 33, A. Ghose, and S. Tsutsui, Eds. New York, NY, USA: Springer-Verlag New York, 2003, pp. 819–845, doi:10.1007/978-3-642-18965-4_33.

[16] K. Fukunaga, *Introduction to Statistical Pattern Recognition*. 2nd ed., New York, NY: Academic Press Prof., 1990.

[17] P. Fränti, J. Kivijärvi, T. Kaukoranta, and O. Nevalainen, "Genetic Algorithms for Large-Scale Clustering Problems, *The Computer Journal*, vol. 40, no. 9, Sep. 1997, pp. 547-554, doi:10.1093/comjnl/40.9.547.

[18] G. Gan, J. Wu, and Z. Yang, "A genetic fuzzy K-Modes algorithm for clustering categorical data," *Expert Systems with Applications*, vol. 36, no. 2, Mar. 2009, pp. 1615-1620, doi:10.1016/j.eswa.2007.11.045.

[19] D. E. Goldberg, *Genetic Algorithms for Search, Optimization, and Machine Learning*. 1st ed., Reading, MA: Addison-Wesley Longman, 1989.

[20] L. O. Hall, I. B. Özyurt, and J. C. Bezdek, "Clustering with a genetically optimized approach," *IEEE Trans. Evol. Comput.* vol. 3, no. 2, Jul. 1999, pp. 103–112, doi:10.1109/4235.771164.

[21] J. Han, and M. Kamber, *Data Mining: Concepts and Techniques*. San Francisco, CA, USA: Morgan Kaufmann, 2005.

[22] J. Handl, and J. Knowles, "Evolutionary multiobjective clustering," in *Proc. 8th Int. Conf. Parallel Problem Solving From Nature*

(PPSN VIII), Birmingham, UK, 2004, pp. 1081–1091. X. Yao, E. K. Burke, J. A. Lozano, J. Smith, J. J. Merelo Guervós, J. A. Bullinaria, J. E. Rowe, P. Tiňo, A. Kabán, and H. Schwefel, Eds., in *Lecture Notes in Computer Science*, chapter 109, vol. 3242, pp. 1081–1091, Berlin, Heidelberg: Springer Berlin Heidelberg, 2004, doi:10.1007/978-3-540-30217-9_109.

[23] J. Handl, and J. Knowles, "Exploiting the Trade-off - The Benefits of Multiple Objectives in Data Clustering," in *Proc. of 3rd Int. Conf. on Evolutionary Multi-Criterion Optimization (EMO 2005)*, Guanajuato, Mexico, 2005, pp. 547–560. C. A. Coello Coello, A. H. Aguirre, and E. ZitzlerEds., in *Lecture Notes in Computer Science*, chapter 38, vol. 3410, pp. 547–560, Berlin, Heidelberg: Springer Berlin Heidelberg, 2005, doi:10.1007/978-3-540-31880-4_38.

[24] J. Handl, and J. Knowles, "Multi-objective clustering and cluster validation," in *Multi-objective Machine Learning*, Y. Jin, Ed., in *Studies in Computational Intelligence*, chapter 2, vol. 16, pp. 21–48, Berlin, Heidelberg: Springer Berlin Heidelberg, 2006, doi:10.1007/3-540-33019-4_2.

[25] J. Handl, and J. Knowles, "An evolutionary approach to multiobjective clustering," *IEEE Trans. Evol. Computat.*, vol. 11, no. 1, Feb. 2007, pp. 56-76, doi:10.1109/TEVC.2006.877146.

[26] S. Hettich, C. Blake, and C. Merz, "UCI repository of machine learning databases," 1998. [Online]. Available: http://www.ics.uci.edu/ mlearn/MLRepository.html

[27] E. R. Hruschka, R. J. G. B. Campello, A. A. Freitas, and A. C. P. L. F. de Carvalho, "A survey of evolutionary algorithms for clustering," *IEEE Trans Syst Man Cybern C*, vol. 39, no. 2, Mar. 2009, pp. 133–155, doi:10.1109/TSMCC.2008.2007252.

[28] Z. Huang, "Clustering large data sets with mixed numeric and categorical values," in *Proc. 1st Pac. Asia Knowl. Discovery Data Mining Conf.*, World Scientific, Singapore, 1997, pp. 21–34.

[29] Z. Huang, "Extensions to the k-means algorithm for clustering large data sets with categorical values," *Data Mining Knowl. Discovery*, vol. 2, no. 3, 1998, pp. 283-304.

[30] Z. Huang, and M. K. Ng, "A fuzzy k-modes algorithm for clustering categorical data," *IEEE Trans. Fuzzy Syst.*, vol. 7, no. 4, Aug. 1999, pp. 446-452, doi:10.1109/91.784206.

[31] Z. Huang, M. K. Ng, H. Rong, and Z. Li "Automated variable weighting in k-Means type clustering," *IEEE Trans. Pattern Anal. Machine Intell.*, vol. 27, no. 5, Oct. 2005, pp. 657-668.

[32] A. K. Jain, M. N. Murty, and P. J. Flynn, "Data clustering: a review," *ACM Comput. Surveys.*, vol. 31, no. 3, Sep. 1999, pp. 264–323, doi:10.1145/331499.331504.

[33] L. Jie, G. Xinbo, and J. Li-cheng, "A GA-based clustering algorithm for large data sets with mixed numeric and categorical values," in *Proc. of the 5th Int. Conf. on Computational Intelligence and Multimedia Applications (ICCIMA 2003)*, IEEE press, Xi'an, China, Sep. 2003, pp. 102–107, doi: 10.1109/ICCIMA.2003.1238108.

[34] H. Jutler, "Liniejnaja modiel z nieskolmini celevymi funkcjami (linear model with several objective functions)," *Ekonomika i Matematiceckije Metody, (In Polish)*, vol. 3, no. 3, pp. 397–406, 1967.

[35] L. Kaufman, and P. J. Rousseeuw, *Finding Groups in Data: An Introduction to Cluster Analysis*. New York: John Wiley & Sons, 1990.

[36] J. Kleinberg, "An impossibility theorem for clustering," in *Advances in Neural Information Processing Systems*, vol. 15, S. Becker, S. Thrum, and K. Obermayer, Eds., Cambridge, MA: MIT Press, 2002, pp. 446–453.

[37] E. E. Korkmaz, J. Du, R. Alhajj, and K. Barker, "Combining advantages of new chromosome representation scheme and multiobjective genetic algorithms for better clustering," *Intell. Data Anal.*, vol. 10, no. 2, Mar. 2006, pp. 163-182.

[38] M. H. C. Law, A. P. Topchy, and A. K. Jain, "Multiobjective data clustering," in *Proc. IEEE Comput. Soc. Conf. on Comput. Vision and Pattern Recognit (CVPR 2004)*, vol. 2, IEEE press, Washington, DC, Jun.-Jul. 2004, pp. 424–430, doi:10.1109/CVPR.2004.1315194.

[39] U. Maulik, and S. Bandyopadhyay, "Genetic algorithm-based clustering technique," *Pattern Recognit.*, vol. 3, no. 9, Sep. 2000, pp. 1455–1465.

[40] U. Maulik, S. Bandyopadhyay, and I. Saha, "Integrating clustering and supervised learning for categorical data analysis," *IEEE Trans. Syst. Man. Cybern. A*, vol. 40, no. 4, Jul. 2010, pp. 664-675, doi:10.1109/TSMCA.2010.2041225.

[41] P. Merz, and A. Zell, "Clustering gene expression profiles with memetic algorithms," in *Proc. of the 7th Int. Conf. on Parallel Prob. Solving from Nature (PPSN VII)*., Lecture Notes in Computer Science, chapter 78, vol. 2439. Berlin, Heidelberg: Springer Berlin Heidelberg, 2002, pp. 811-820, doi:10.1007/3-540-45712-7_78.

[42] A. Mukhopadhyay, U. Maulik, and S. Bandyopadhyay, "Multiobjective genetic fuzzy clustering of categorical attributes," in *Proc. 10th Int. Conf. on Inf. Technol. (ICIT 2007)*, IEEE press, Orissa, India, Dec. 2007, pp. 74-79, doi:10.1109/ICIT.2007.13.

[43] A. Mukhopadhyay, U. Maulik and S. Bandyopadhyay, "Multiobjective genetic algorithm-based fuzzy clustering of categorical attributes," *IEEE Trans. Evol. Computat.*, vol. 13, no. 5, Oct. 2009, pp. 991–1005, doi:10.1109/TEVC.2009.2012163.

[44] H. Pan, J. Zhu, and D. Han, "Genetic algorithms applied to multiclass clustering for gene expression data," *Genomics, Proteomics, Bioinformatics*, vol. 1, no. 4, Nov. 2003, pp. 279–287.

[45] V. Pareto, *Manuale di Economia Politica*. Milan:Piccola Biblioteca Scientifica, 1906. Translated into English by A. S. Schwier, and A. N. Page, Manual of Political Economy, London:Kelley Publishers, 1971.

[46] L. Parsons, E. Haque, and H. Liu, "Subspace clustering for high dimensional data, a review," *SIGKDD Explor. Newsl.*, vol. 6, no. 1, Jun. 2004, pp. 90–105, doi:10.1145/1007730.1007731.

[47] S. Lloyd, "Least squares quantization in PCM," *IEEE Trans. Inf. Theor.*, vol. IT-28, no. 2, Mar. 1982, pp. 129–137, doi: 10.1109/TIT.1982.1056489.(Original version: Technical Report, Bell Labs, 1957).

[48] J. B. MacQueen, "Some methods for classification and analysis of multivariate observations," in *Proc. 5th Berkeley Symp. Mathematical Statistics and Probability*, Statistical Laboratory of the University of California, Berkeley, 1967, pages 281–297.

[49] J. D. Schaffer, "Multiple objective optimization with vector evaluated genetic algorithms," in *Proc. 1st Int. Conf. Genetic Algorithms*, L. Erlbaum Associates, Pittsburgh, PA, Jul. 1985, pp. 93–100.

[50] P. Scheunders, "A genetic c-means clustering algorithm applied to color image quantization," *Pattern Recognition*, vol. 30, no. 6, Jun. 1997, pp. 859-866, doi:10.1016/S0031-3203(96)00131-8.

[51] Z. Zheng, M. Gong, J. Ma, L. Jiao, and Q. Wu, "Unsupervised evolutionary clustering algorithm for mixed type data," in *Congress of Evolutionary Computation (CEC 2010)*, Barcelona, Spain, Jul. 2010, pages 1–8, doi:10.1109/CEC.2010.5586136.