Journal of Advances in Computer Research
Quarterly pISSN: 2345-606x eISSN: 2345-6078
Sari Branch, Islamic Azad University, Sari, I.R.Iran
(Vol. 8, No. 3, August 2017), Pages: 69-77
www.jacr.iausari.ac.ir

Feature Selection and Clustering By Multi-Objective Optimization

Seyedeh Mohtaram Daryabari[1], Farhad Ramezani ✉[2]

1) Department of Computer Engineer, Rouzbahan Institute, Sari, I. R. Iran
2) Department of Computer Engineering, Sari Branch, Islamic Azad University, Sari, I. R. Iran

s.mitra.daryabari@gmail.com; ramezani.farhad@iausari.ac.ir

Received: 2016/07/13; Accepted: 2016/09/17

Abstract

In this paper, feature selection and clustering is formulated simultaneously by using evolutional multi-objective algorithm. Archived multi-objective NSGA-II is hybridized with k-medoids algorithm to use global searching capabilities of GA with local searching capabilities of k-medoids for suitable centers of clusters and selecting suitable subset of features identifying the correct partitioning. Number of clusters should be determined as an input parameter by user. After determining number of clusters, archive string be generate randomly. In every solution of archived, center of clusters and features is determined. Objective functions are inter-cluster distance, intra-cluster distance and number of feature selection. Three objective functions are optimized simultaneously for partitioning and feature selection. Crossover and mutation operators are modified to solve the problem. In order to selecting final solution from pare to front, are modified to solve the problem is calculated. The Proposed algorithm were compared with other three clustering algorithms on seven UCI standard datasets and could improve results averagely 0.09 percent compared to FeaClusMoo, 0.28 percent compared to VGAPS-Clustering and 0.49 percent compared to K-means.

Keywords: *Clustering; Data Mining; Feature Selection; Multi-objective Optimization; NSGA-II*

1. Introduction

Generally, features are characterized as: (i) Relevant: features which have an influence on the output and their role cannot be assumed by the rest, (ii) Irrelevant: features not having any influence on the output, (iii) Redundant: a feature can take the role of another [1]. In the past years in the applications of machine learning or pattern recognition, the domains of features have expanded from tens to hundreds of variables or features used in those applications [2]. Increasing of dimensionality of data in many methods of feature selection and feature extraction with regard to efficiency are a hard challenge and simplest way in order to decreasing dimensions is feature selection [3].

A feature selection framework generally consists of two parts: a searching engine used to determine the promising feature subset candidates, and a criterion used to determine the best candidate [1]. Feature Selection (variable elimination) helps in understanding data, reducing computation requirement, reducing the effect of curse of dimensionality and improving the predictor performance [4]. The optimality of a feature subset is measured by an evaluation criterion [5].

Based on whether the training data is labeled or not, feature selection methods are broadly divided into supervised methods [6], [7], [8] and unsupervised methods [9], [10], [11]. The depth of treatment of various subjects reflects the proportion of papers covering them: the problem of supervised learning is treated more extensively than that of unsupervised learning [12]. As most data is unlabeled, unsupervised feature selection attracts increasing attention in recent years [13].

One of important application in machine learning is unsupervised data clustering. Clustering is a division of data into groups of similar objects and each group, by name of cluster, includes some objects that are similar to each other and objective from groups are different [14]. For Unsupervised Feature Selection, definition relationship of feature due to lack of label information is very difficult and some alternative criteria have been proposed such as data variance, data similarity and data separability [13]. Since using validity for various kind of dataset cannot work same, it is natural to use multi-objective simultaneously in order to obtaining various features of data [15].

Evolutionary techniques for multi-objective optimization are currently gaining significant attentions from researchers in various fields due to their effectiveness and robustness in searching for a set of global trade-off solutions [5]. Use multi-objective clustering techniques that optimize more than one cluster validity index simultaneously, is useful for clustering and, leading to high-quality results [15]. Among multi-objective algorithm can be noted Pareto archived evolution strategy (PAES) [16], strength Pareto evolutionary algorithm-2 (SPEA2) [17], Non-dominated sorting genetic algorithm-II (NSGA-II) [18].

In recent years, feature selection and clustering method have been introduced simultaneously by using multi-objective algorithm. Multi-objective genetic algorithm based k-clustering have been used for feature selection and clustering simultaneously [19], [20]. In [21] a multi-objective method of feature selection for unsupervised clustering is presented. In [22] for protection of similar samples a multi-objective optimization is used and the feature space is reduced. In [23] simulated annealing optimization method along with an archive for selection and clustering features are used simultaneously. In this paper, number of clusters is known and NSGA-II is utilized as the background optimization methodology. The improvements of proposed method compared to the other three algorithms are shown in Table 2.

The rest of this paper has been organized as follow: In section 2, a definition of multi-objective optimization and non-dominated sorting genetic algorithm II has been presented. Thereafter, in Section 3 describes the newly proposed method in detail. At section 4, the experimental results have been report. Finally section 5 summarizes the work with concluding remarks.

2. Multi-objective Optimization

If it is assumed that there are non-commensurable k objectives and none of objectives have priority than other objectives. Generally, multi-objective optimization [24] can be defined as the problem of optimizing a vector of non-commensurable which objectives compare to each other. This can lead to finding a parametric set m for $\min f(m)$ that $m \in £$. $m = \{m_1, \dots, m_n\}$ is an individual vector with n parameter and $£$ is set individual vectors. $\{f_1, \dots, f_k\}$, are k objectives to be minimized. There are sets of solutions in multi-objective problems that known as Pareto optimal set. Since, genetic algorithms (GAs) work with a population of points, it seems natural that GAs have been

used in multi-objective optimization problem in order to obtaining numbers of solution simultaneously [24].

2.1 Non-dominated Sorting Genetic Algorithm-II (NSGA-II)

In Non-dominated sorting genetic algorithm-II (NSGA-II) [18], for the first time density estimation metric has been defined and then crowded-comparison operator has been introduced. In density estimation in order to getting an estimate of the density of solutions around a special solution in the population, have been calculated the average distance of two points on either side of this point along each of the objectives. Crowding-distance value is obtained from sum of related individual distance values for each objective. The smaller the crowding distance is, the more the crowding around the solution. Crowded-comparison operator ($\prec_n$) has been used for guiding selection process in algorithm. It is assumed that every individual in the population has two attributes: non-domination rank (i_{rank}) and crowding distance $(i_{distance})$ and be defined a partial order $\prec_n$ as:

$$i \prec_n j \quad if\ (i_{rank} < j_{rank}) \quad (1)$$
$$or\ ((i_{rank} = j_{rank})$$
$$and\ (i_{rank} > j_{rank}))$$

Between two solutions with differing non-domination ranks, lower solution has been chosen and if two solution belong to a non-domination rank, then solution in low density be selected. NSGA-II procedure has been shown in Figure 1.

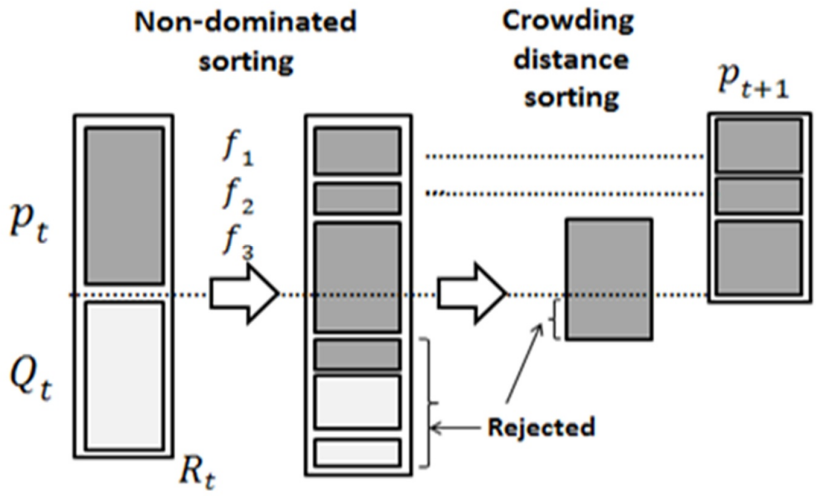

Figure 1. NSGA-II procedure

In tth generation, p_t is parent population and Q_t is children population and $R_t = p_t \cup Q_t$. f_1 Is the best non-domination rank and f_2 is the next best level and etc. size of p_{t+1} and p_t are N and size of R_t is$2N$.

3. Proposed Method

In this method, number of clusters should be determined as an input parameter by user. Archive multi-objective NSGA-II algorithm has been used as a background

optimization strategy. After determining number of clusters, archive string generated randomly. Every individual in archive include two parts. The first part include k number of randomly points that selected from dataset which is the center of the clusters and the second part include randomly binary string from $\{0, 1\}$ along number of dataset points which specifies the composition of features. An example of a string is given below.

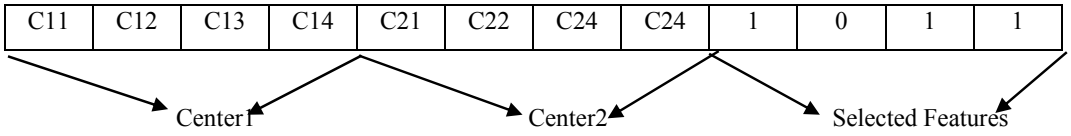

Figure 2. Example of cluster centers and selected features in the archive

According to Fig.2 it is assumed that string in the archive contains two clusters, {C11, C12, C13, C14} and {C11, C12, C13, C14}, and only the first, third and fourth features have been considered. After selecting the string, the Dataset is allocated to selected centers in the archive strings using [16] K-medoids. Its steps are as follows:

1- Euclidean distance of each point like x in the dataset calculated by selected centers in the solution

2- Each point is assigned for a center that has lowest Euclidean distance to it.

3- Average distance within each cluster is calculated based on equation 2.

$$M = \frac{1}{m} \sum_{i=1}^{m} (xi - cj)^2 \qquad \text{j} \in 1 \ldots \text{k} \tag{2}$$

Where m is number of cluster in jth cluster.

4- Nearest data to calculated value, is chosen as center of cluster.

5- Again points are assigned to new centers. This is repeated so that the rest allocated points to any cluster does not change.

Only those feature which 1 of value in that specific string have been considered.

Three objective functions have been defined for optimization. The first validity index of inter-cluster distance is obtained based on equation 3.

$$\text{f1} = \sum_{i=1}^{k} \sum_{j=1}^{m} \sqrt{(xj - ci)^2} \tag{3}$$

Where c_i is center of cluster ith and x_j is jth point in the same cluster. The second validity index of intra-cluster distance is calculated based on equation 4.

$$\text{f2} = \sum_{i=1}^{k} \sum_{j=i+1}^{k} \sqrt{(ci - cj)^2} \tag{4}$$

Where ci is center of cluster ith and cj is center of cluster jth. The third objective function is number of selected features in considered solution. This is obtained according to the equation 5.

$$f3 = \sum_{i=1}^{d} f_i \quad (5)$$

Where d is the number of dimensions and f_i , is the i_{th} features in binary string with the value of {0, 1}. The purpose is decreasing inter-cluster distance and number of features and increasing intra-cluster distance. As a result optimized functions related to a particular string, is obtained by equation 6.

$$obj = \left\{f_1, {}^{1}/_{f_2}, f_3\right\} \quad (6)$$

In calculating objective functions, only those feature which 1 of value in that specific string have been considered. NSGA-II algorithm minimizes the three objective functions.

With regard to solutions, mutation is this method has two parts. The first part is related to centers and the second part belongs to binary string.

1- In the center part, one of them has been chosen randomly and then that center has been replaced with a randomly selected point from dataset.

2- In binary part, first two bit is selected randomly and values 0, 1 are changed to each other. If two selected bits both were 0 or 1, one of these two bits be chosen randomly and is inversed.

For selecting the best solution and determining similarity of obtained solution, Minkowski Score [25] value is calculated, that is base equation 7.

$$MS(T,S) = \sqrt{\frac{n_{01}+ n_{10}}{n_{11}+ n_{10}}} \quad (7)$$

Where T is true solution and S is obtained solution. n_{11} is pair number of elements, that in an equal cluster are both T, S. n_{01} is pair number that are in an equal cluster in S and n_{10} is pair number that in are in an equal cluster in T. After calculated Minkowski Score value for all solution of pareto front, the solution that have the lowest value be can obtained as the best solution.

4. Test and Comparison

In order to evaluating proposed method, standard dataset, UCI, can be used for experimental purposes. The number of generations was considered 100. Data collection and description of its size, the number of dimensions and the number of clusters is shown in Table 1 . Proposed method efficiency has been compared to three algorithms FeaClusMoo, VGAPS-clustering, k-means. Minkowski points of the last partition values obtained by the proposed method and three other algorithms are reported in Table 1. The improvements of proposed method compared to the other three algorithms are shown in Table 2. The abundance charts and estimation of normal distribution fits for all three categories according to features calculated using 2015 MATLAB software by Distribution Fitting Tools toolbar, the example of Iris data collection is shown in

Figure 3. After normalize the data of each feature between the range of zero to one, covariance and mean of the individual characteristics of each category is calculated.

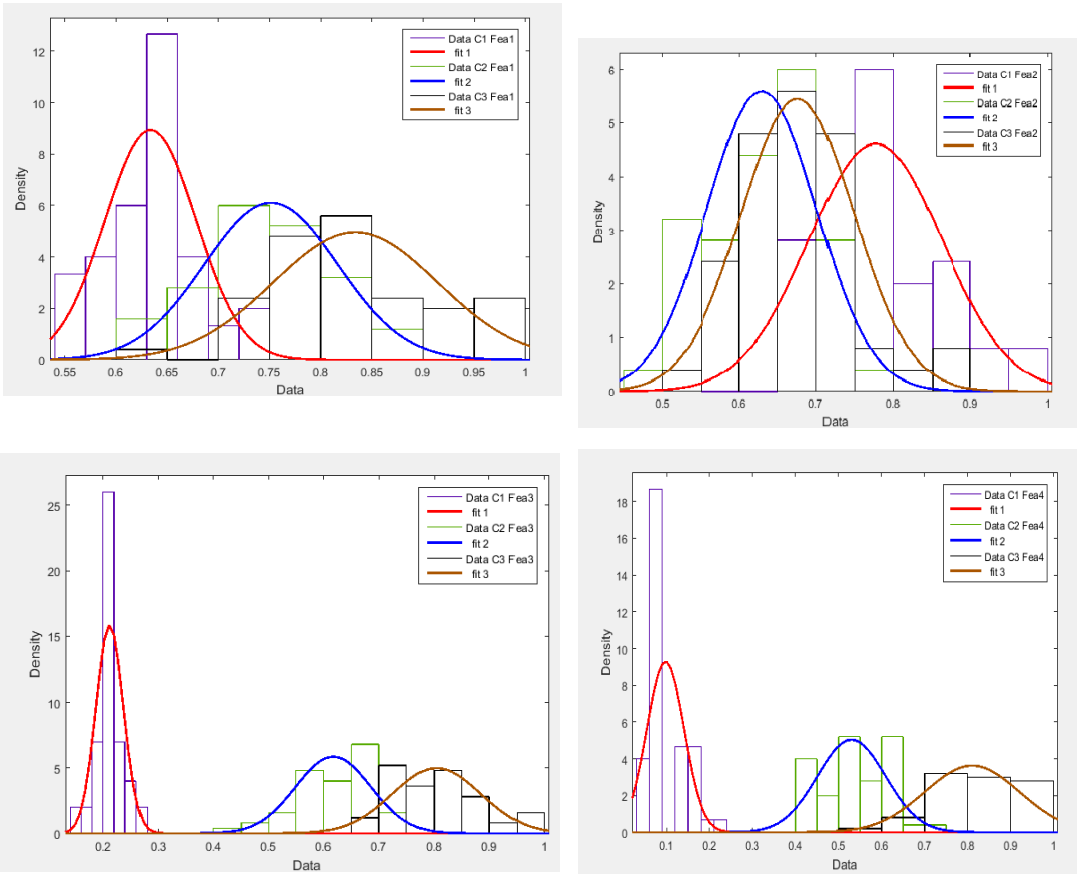

Figure 3. abundance charts and average normal distribution base on group separation for dataset iris

Values of Minkowski Score in relation to final partition that is obtained by 3 algorithms have been reported in Table 1.

Table 1. Results on different datasets, UCI by means of FeaClusMoo, VGAPS-Clustering, K-means algorithms and proposed method

			FeaClusMoo		VGAPS-Clustering		k-means		Proposed method	
Dataset	**N**	**d**	**F**	**MS**	**F**	**MS**	**F**	**MS**	**F**	**MS**
Iris	150	4	3,4	0.44	All	0.62	All	0.68	4	0.38
Cancer	683	9	1-6	0.31	All	0.37	All	0.37	1,3,5,6,8	0.36
Newthy	215	5	1,2,4,5	0.54	All	0.58	All	0.94	2,4,5	0.53
Wine	178	13	1,6,7	0.67	All	1.12	All	1.40	1,7	0.58
LiverDis	345	6	1,2,5	0.98	All	0.98	All	0.98		0.97
LungCan	33	56	1-4,7,8 10,11,13	0.70	All	1.24	All	1.45	2,4,5,9 14-16	0.48

			16,19-23						18,23,25		
			25-27						28,32,33		
			29-31						37,38,41		
			32-39						43-45		
			42-45						47		
			47-49						51-55		
			51-53								
Glass	214	10	1-5	1.05	All	1.10	All	1.69	1,6,7,8	0.72	

Where N is number of data points and d is original number of features. MS is Minkowski Score value and F is number of selected features.

Table 2: The improvements of proposed method compared to the FeaClusMoo, VGAPS-Clustering, K-means

Dataset	***FeaClusMoo***	***VGAPS-Clustering***	***K-means***
Iris	0.06	0.24	0.3
Cancer	-0.05	0.01	0.01
Newthy	0.01	0.05	0.41
Wine	0.09	0.54	0.82
LiverDis	0.01	0.01	0.01
LungCan	0.22	0.76	0.97
Glass	0.33	0.38	0.97

In order to feature selection and clustering simultaneously, archives NSGA-II have been presented. Number cluster can be determined by user. Three objective functions: inter-cluster, intra-cluster distance and decreasing number of features have been selected for optimization.

As the results in Table 1 and Table 2 show, the advantage of the proposed method is well clustering of decimal data collection, however, clustering the data set with the correct value and low dimensions cannot perform well. As it was reported for Cancer data set with the correct values and low dimensional data sets the FeaClusMoo algorithm is more efficient than the proposed method, while the proposed algorithm for data collection, such as LungCan data collection with integer values and high dimensions, perform better than other methods.

5. Conclusions and Future Works

In this paper, number of clusters can be determined by user. Archive multi-objective NSGA-II, feature selection, the appropriate cluster centers selection and true partitioning is determined evolutionary. Assigning dataset points to selected centers in each string has been done by using k-medoids algorithm. The first objective function

has been considered inter-cluster density and second objective has been considered intra-cluster density and third objective function has been considered number of features. Three objective functions have been optimized simultaneously for partitioning and feature selection. Crossover and mutation operators are modified to solve the problem. In order to selecting final solution from pareto front, are modified to solve the problem is calculated. Determined clusters and number of features from proposed method by FeaClusMoo, VGAPS-clustering and k-means method for seven datasets have been demonstrated in Table 1. The improvements of proposed method compared to the other three algorithms are shown in Table 2 and show that Minkowski Score value from proposed method is less than three other algorithms.

In this paper, number of clusters is determined as parameter by user. It is suggested that for optimizing, it is better to presenting a way that be able to use suitable number of clusters automatically and also be able to work well for datasets with integer numbers and low dimensions.

References

[1] S. B. Kotsiantis, "Feature selection for machine learning classification problems: a recent overview," Artificial Intelligence Review, 2011.

[2] M. Anirban, M. Ujjwal and B. Sanghamitra, "A Survey of Multiobjective Evolutionary Algorithms for Data Mining: Part I," IEEE TRANSACTIONS ON EVOLUTIONARY COMPUTATION, vol. 18, no. 1, pp. 4 - 19, FEBRUARY 2014.

[3] K. Samina, K. Tehmina and N. Shamila, "A Survey of Feature Selection and Feature Extraction Techniques in Machine Learning," in *Science and Information Conference*, London, 2014.

[4] L. Huan and L. Yu, "Toward integrating feature selection algorithms for classification and clustering," IEEE Transactions on Knowledge and Data Engineering, vol. 17, no. 4, pp. 491 - 502, April 2005.

[5] K. C. Tan, T. H. Lee and E. F. Khor, "Evolutionary Algorithms for Multi-Objective Optimization: Performance Assessments and Comparisons," Artificial Intelligence Review, vol. 17, no. 4, p. 251–290, June 2002.

[6] S. Le, S. Alex and G. Arthur, "Supervised Feature Selection via Dependence Estimation," in ICML '07 Proceedings of the 24th international conference on Machine learning , New York, 2007.

[7] M. Dash and H. Liu, "Feature Selection for Classification," Intelligent Data Analysis, vol. 1, no. 3, pp. 131-156, 1997.

[8] K. Nojun and C. Chong-Ho, "Input Feature Selection for Classification Problems," IEEE TRANSACTIONS ON NEURAL NETWORKS, vol. 13, no. 1, pp. 143-159, JANUARY 2002.

[9] P.Mitra, C.A.Murthy and S.K.Pal, "Unsupervised feature selection using feature similarity," IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 24, no. 3, pp. 301 - 312, Mar 2002.

[10] J. G. Dy and C. E. Brodley, "Feature Selection for Unsupervised Learning," pp. 845--889, 2004.

[11] D. Cai, Z. Chiyuan and X. He, "Unsupervised Feature Selection for Multi-Cluster Data," in ACM, New York, 2010.

[12] G. Isabelle and A. Elisseeff, "An Introduction to Variable and Feature Selection," Journal of machine learning research, pp. 1157-1182, 2003.

[13] J. Tang, X. Hu, H. Gao and H. Liu, "Discriminant Analysis for Unsupervised Feature Selection," in *SIAM International Conference on Data Mining*, 2014.

[14] "A Survey of Clustering Data Mining Techniques," in Grouping Multidimensional Data , Berlin , Springer Berlin Heidelberg, 2006, pp. 25-71.

[15] A. Mukhopadhyay, U. Maulik and S. Bandyopadhyay, "Survey of Multiobjective Evolutionary Algorithms for Data Mining: Part II," IEEE TRANSACTIONS ON EVOLUTIONARY COMPUTATION, vol. 18, no. 1, pp. 20 - 35, FEBRUARY 2014.

[16] J. Knowles and D. Corne, "The Pareto archived evolution strategy: a new baseline algorithm for Pareto multiobjective optimisation," in Evolutionary Computation, 1999. CEC 99. Proceedings of the 1999 Congress on, August 2002.

[17] E. Zitzler, M. Laumanns and L. Thiele, "SPEA2: Improving the Strength Pareto," Eurogen, 2001.

[18] "A fast and elitist multiobjective genetic algorithm: NSGA-II," IEEE Transactions on Evolutionary Computation, vol. 6, no. 2, pp. 182 - 197 , 2002.

[19] D. Dutta, P. Dutta and J. Sil, "Categorical Feature Reduction Using Multi Objective Genetic Algorithm in Cluster Analysis," in Transactions on Computational Science XXI , Berlin, Springer Berlin Heidelberg, 2013, pp. 164-189.

[20] D. Dutta, P. Dutta and J. Sil, "Simultaneous Continuous Feature Selection and K Clustering by Multi Objective Genetic Algorithm," in Advance Computing Conference (IACC), Feb. 2013 .

[21] E. d. l. Hoz, E. d. l. Hoz and A. Ortiz, "Feature selection by multi-objective optimisation: Application to network anomaly detection by hierarchical self-organising maps," Knowledge-Based Systems, vol. Volume 71, p. 322–338, November 2014.

[22] P. P. Kundu and S. Mitra, "Multi-objective optimization of shared nearest neighbor similarity for feature selection," Applied Soft Computing, vol. 37, p. 751–762, December 2015.

[23] S. Saha, R. Spandana, A. Ekbal and S. Bandyopadhyay, "Simultaneous feature selection and symmetry based clustering usingmultiobjective framework," Applied Soft Computing, vol. 29, p. 479–486, April 2015.

[24] N. Siinivas and K. Deb, "Muiltiobj ective Optimization Using Nondominated Sorting in Genetic Algorithms," the Massachusetts Institute of Technology, vol. 2, no. 3, pp. 221-248, 1994.

[25] S. Bandyopadhyay and S. Saha, "A Point Symmetry-Based Clustering Technique for Automatic Evolution of Clusters," IEEE Transactions on Knowledge and Data Engineering, vol. 20, no. 11, pp. 1441 - 1457, Nov. 2008.