Available online at www.sciencedirect.com

ScienceDirect

Procedia Computer Science 57 (2015) 1074 – 1080

Procedia Computer Science

3rd International Conference on Recent Trends in Computing 2015 (ICRTC-2015)

Feature Selection using K-Means Genetic Algorithm for Multi-objective Optimization

M.Anusha[a*], Dr. J.G.R.Sathiaseelan[b]

[a]*Department of Computer Science, Bishop Heber College,Trichy.TN.INDIA.*
[b]*Department of Computer Science, Bishop Heber College,Trichy.TN.INDIA.*

Abstract

Multi-objective genetic-clustering algorithms are based on optimization which optimizes several objectives simultaneously. In multi-objective optimization problem (MOP), different objective function may have different properties. In the previous paper, multi-objective optimization on neighbourhood learning using k-means genetic algorithm (NLMOGA), was proposed and applied to several real-life data sets. This research paper aims to extend NLMOGA by maximizing the compactness and the accuracy of the solution through constraint feature selection on the selected sub-population. A new population is generated using NLMOGA and a constrained feature selection is applied to each sub-population. This method is developed to determine a suitable or closest set of objects from the group objects which will improve the robustness of NLMOGA for different instances of MOPs. In NLMOGA, a solution is selected from global population repository and then neighbourhood learning is made to promote the evolution of each objective for the selected solution. The effectiveness of this approach is evaluated with various real-life benchmark gene expression data sets.

Peer-review under responsibility of organizing committee of the 3rd International Conference on Recent Trends in Computing 2015 (ICRTC-2015)

Keywords: multiobjective optimization; feature selection; genetic algorithms; clustering.

1. Introduction

Clustering is a data mining technique and it is widely used to group the similar set of data. The multi-objective

* Corresponding author. Tel.: +91-82-20-376755;
E-mail address: anusha260505@gmail.com

Peer-review under responsibility of organizing committee of the 3rd International Conference on Recent Trends in Computing 2015 (ICRTC-2015)
doi:10.1016/j.procs.2015.07.387

optimization problems have two or more objectives which are subjected to certain constraints that need to be optimized simultaneously. The multi-objective clustering methods are intended on grouping data using multiple benchmarks problems [1]. In general, a genetic algorithm for a multi-objective optimization problem, we need to initialize the random global population of solutions having n real valued variables and the population is optimized to acquire new population by the following operators namely, selection, genetic operators and elitism. Selection is carried out through certain condition that is solution with greater suitability is chosen for the intermediate mating pool. The genetic operators like crossover and mutation is used to generate a new improved subpopulation. Elitism is applied to certain gene that does not require any further modifications. The process ends at certain terminating condition. The multi-objective clustering technique uses genetic algorithm [2] to generate clusters which are simultaneously proceed by multiple objectives to obtain a collection of non-dominated solutions with several trade-off among the calculated objectives. Many real world issues occur in various fields such as data mining [3], artificial neural network [4] and so on. The most frequently approach is a-posteriori approach where the preference information is set after the Pareto optimal solution is obtained through genetic algorithm. The solution obtained above will be more scattered. In order to maintain the compactness of the solution from the diverse set of non-dominated solution, very closest front to the true non-dominated solution is found. Based on the dominated front a trade-off solution will be generated. From the trade-off solution a single object is chosen by using fitness information [5]. In case of a-priori approach preference information or criteria is included in prior to process genetic algorithm [6].

Recently, high-dimensionality problems with regard to number of features have gained increasing importance. The problem feature selection lies in the determination of an optimal feature subset among the full set of features. This measure can optimize the adopted criteria for clustering accuracy. Feature selection helps in understanding data, reducing computation requirement and improve the performance of algorithm. The use of optimization algorithms like genetic algorithm in feature selection has gained attention over traditional methods [7]. The feature selection can be defined as the process of choosing a minimum subset of r features from global population repository of m features($r<m$) which in turn reduces the space with high cluster accuracy having the only the selected features as its group. A feature selection algorithm includes four steps for its process. They are subset generation, subset evaluation, stopping criterion and result validation [8]. Subset generation is a searching process which generates subsets of features for evaluation. Each generated subset is evaluated by some specific fitness measure and compared with the previous best one with respect to this objective measure. If the new subset is found better than the previous best subset is replaced by new subset. The goal of this research paper is to find Pareto optimal set based on the feature selection problem using EKMGA [9]. The applicability and effectiveness of the proposed approach is tested by conducting experiments using several real-life benchmark data sets.

The remainder of the paper is organized as follows. Section 2 briefly summarizes the related work on feature selection for multi-objective evolutionary clustering. Section 3 proposed algorithm for feature selection problem. Section 4 describes the experimentation and discuss the result. Finally, Section 5 ends with conclusion with future work.

2. Literature Survey

Kung-jeng wang et al. [10] proposed a novel method for feature selection using opposite sign test (OST) for an improved artificial immune recognition system. However the algorithm needs improvement in setting testing parameter. Girish chandrashekar et al. [11] presented a massive study on various feature selection methods with variable limitation on filter, wrapper and embedded methods. Chih-fong tsai et al. [12] discussed feature selection using priorities which lacks in accuracy with high computational cost. Hu xia et al. [13] developed a subspace clustering with a multi-objective evolutionary approach is addressed with feature weight learning method for selecting features in a high-dimensional data. However, the algorithm fails to identify the clusters with poor time complexity.

Nibaran das et al. [14] identified the feature with the local regions in the given pattern to evaluate the handwritten Bangla digits. The performance of this method is poor in identifying the charaters. Pedrycz et al. [15] introduced a feature space reduction guided by structure retention criterion. However the algorithm fails attain cluster compactness with lack of accuracy. Emel kizilkaya aydogam et al. [16] proposed a hybrid genetic algorithm using fuzzy rule for classifying high dimensional problem. In this method the author utilizes an inter-programming formulation (IPF) for selecting the features in the algorithm. The effectives of the algorithm in terms of setting the rules need to be

compressed for better feature selection. Ratta et al. [17] stated an improved feature selection for real time disruption prediction on Joint European Torus (JET) with the help of advanced predictor of disruption (APODIS) feature selector. The performance of the approach has to be improved for better results.

3. Proposed Algorithm

3.1. The objectives of FS-NLMOGA

The proposed algorithm uses three objective functions which simultaneously maximize the inter-cluster distance (diversity) and minimize the intra-cluster distance (compactness) with high accuracy for optimal clustering. The feasible approach to obtain the Pareto front solution is to minimize the inter-cluster distance within the feature set and maximizing the diversity between cluster classes. In-order-to find the more similar cluster objects, a criterion-based feature selection is applied as input. This method eliminates the outliers from the cluster set and generates very close objects to the selected feature. This method also increases the clustering accuracy of the Pareto front solutions. Therefore, the proposed method simultaneously optimizes three objectives by maximizing the diversity and accuracy and minimizing the compactness of the cluster solution.

3.2. Feature Selection and Fitness Evaluation

The proposed feature selection algorithm adopts NAGA II as is base. Since the basic structure uses genetic algorithm, we have define the global population. Naturally, the original feature set without the matching class is encoded as m-bit binary string. The value of m represents the key of the original feature set. The feature selection for the particular group is represented as zero or one. The total distance between the selected feature subset and the corresponding cluster class (diversity) is stated below:

$$div(p,q) = \frac{1}{|p|}\sum_{i=1}^{|p|} |\, nl(n,p)\,| \tag{1}$$

Where |p| is the key of the selected feature subset of p and q is the coordinating class which holds the result. *nl* is the neighbour learning algorithm of NLMOGA. The absolute distance between the selected feature and the cluster class is averaged by |p|. The overall distance between the selected feature subset (compactness) is shown below:

$$comp(p) = \frac{1}{SFO}\sum_{a=1}^{|p|}\sum_{b=a++}^{|p|} |\, nl(x_a, x_b)\,| \tag{2}$$

Where SFOis the selected featured objects in the feature subset p. SFO eliminates the noisy data from the group to get more compact cluster solutions. The overall cluster accuracy is validated with the help of Silhouette index.

3.3. Implementation of FS-NLMOGA

The complete implantation of FS-NLMOGA is presented below.

Algorithm FS-NLMOGA

Input: PC (Probability of crossover), PM (Probability of mutation), GPR(Size of global population repository), MGEN (Maximum generations) using EKMGA.

Output: Pareto front solutions.

Begin

1. **While** stopping criterion is not met **do**
2. S←rand_select_population(GPR)
3. **for** x← 1 to p **do**
   (a) Calculate the fitness for objective function using the equation (1) and (2).
   (b) Rank the individual using SFO. Select the closest objects as new_subpopulation.
   (c) Apply neighborhood learning on the selected subpopulation.
   (d) **if** rand[0,1]<PC
   then perform crossover for the particular object
   else apply probability mutation on the selected objects.
   (e) Repeat step (a) until the it reaches the stopping criterion.
4. Return Pareto front solution as result.

3.4. Analysis of Computational Complexity

This paper concenters on the feature selection to select the closest objects in the sub-population. The time taken to find the closest objects in the global population repository using FS-NLMOGA is comparatively lesser than NLMOGA. The time complexity of the proposed algorithm in featured closest objects in selection space is $O(|p|GPR)$.Hence, FS-NLMOGA provides better results in term of efficiency, selecting the closer objects and execution time.

4. Experimental Studies

To evaluate the performance and efficiency of the proposed algorithm NLMOGA, the experiments are conducted using personal computer which uses Windows 7 as operating system. The NLMOGA is implemented using MATLAB 7.0. We analyzed the proposed algorithm using various real life datasets. Cluster validity index called Silhouette index is used to validate the result. Silhouette index value lies between the interval [-1 1]. A value close to 1 means the cluster objects are similar and 0 means dissimilar clusters such that the objects lie far from the clusters, while -1 indicates that the sample are misclassified.

4.1. Data sets

Four real-life data sets are used for experiments. A short description of the data sets in term of size, dimension and number of clusters is provided in Table 1. The real-life data sets are obtained from UCI Machine Learning Repository.

Table 1.Description of Real-life data sets.

Data sets	Size of the data sets	Number of dimensions	Number of clusters
Ionosphere	351	33	2
Iris	150	4	3
Wine	178	13	3
Seed	210	7	3

4.2. Parameter Setting

The number of clusters parameter is fixed for the particular data sets. For the proposed algorithm, the crossover rate is 0.95, mutation rate is 0.01 and population size is 200.

4.3. Comparison of Cluster Accuracy

In order to evaluate the proposed algorithm, it is necessary to define a measure of agreement between two partitions of same data sets. Table 2. shows the results obtained from NLMOGA and FS-NLMOGA. The quality of the cluster is evaluated using Silhoutee index. From the result, it is certain that clustering accuracy of FS-NLMOGA is greater than NLMOGA except the result obtained for ionosphere data set. This is because of

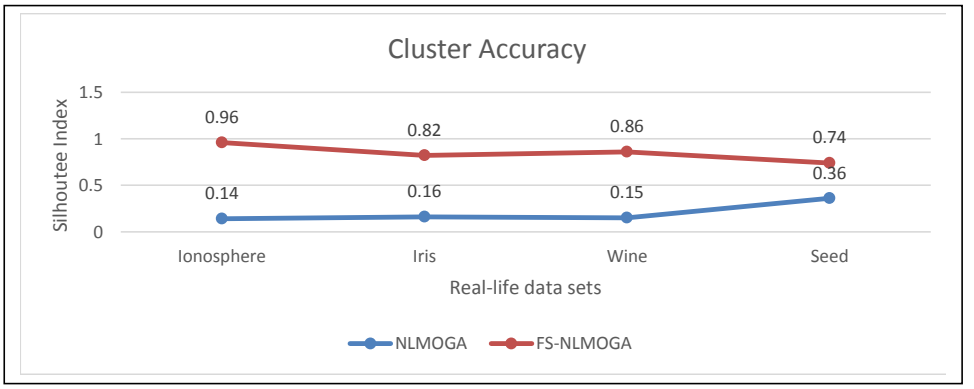

Fig. 1. Cluster Compactness of the real-life data sets

applied in input section. Fig.1. show the cluster accuracy between two algorithms. Hence, we can conclude that FS-MOGA is more efficient than NLMOGA for feature selected clustering problem. The algorithm can also accomplish high dimensional data set.

Table 2. The Results of Clustering Accuracies of NLMOGA, FS-NLMOGA using Silhouette index

Data sets	NLMOGA	FS-NLMOGA
Ionosphere	0.14	0.96
Iris	0.16	0.82
Wine	0.15	0.86

Data sets	NLMOGA	FS-NLMOGA
Seed	0.36	0.74

4.4. Performance Analysis for Cluster Compactness

Table 3. shows the cluster compactness for the four real-life data sets respectively. Since, the proposed Feature Selected-NLMOGA uses criterion-based feature selection, it is proved that proposed algorithm is performing well. Fig.2.showsthe cluster compact between the cluster classes of the real-life data sets.

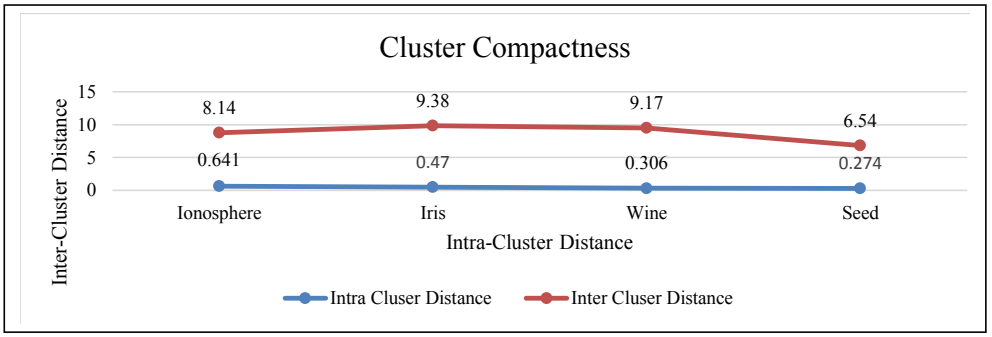

Fig. 2. Cluster Compactness of the real-life data sets

It is inferred that the intra-cluster distance and inter-cluster distance is highly feature selected. Also the proposed algorithm satisfies the objectives by minimizing the intra-cluster distance and maximizing the inter-cluster distance.

Table 3. The Results of Cluster Compactness of FS-NLMOGA

Data sets	Intra-cluster distance with-in cluster	Inter-cluster distance between clusters
Ionosphere	0.641	8.14
Iris	0.470	9.38
Wine	0.306	9.17
Seed	0.274	6.54

5. Conclusion

In this paper, we propose an improved feature selection algorithm using k-means genetic algorithm for multi-objective optimization problem. In contrast to conventional multi-objective genetic algorithm, FS-NLMOGA maximizes two objective functions also minimizes an objective function simultaneously. These functions are optimized simultaneously using feature selected criterion. By using the proposed algorithm, the quality of the cluster is increased. Pareto front solutions contains more compact cluster classes. No weighting parameters are set. The cluster result shows high accuracy than NLMOGA. The performance of the proposed algorithm is tested with several real-life benchmark data sets. The results indicates that the algorithm can simultaneously optimize the chosen objectives by minimizing the intra-cluster distance and maximizing the inter-cluster distance with high accuracy.

While FS-NLMOGA holds great potential on feature selected neighborhood learning, there is a scope to enhance it further by means of constrained crossover on high dimensional data sets. In order to improve the time complexity

and to reduce the computational cost, FS-NLMOGA can be extended with four objectives. We will attempt to address these issues in the future work.

References

1. Anirban Mukhopadhay, Sanghamitra Bandyopadhyay. Survey of Muliobjective Evolutionary Algorithms for Data Mining. *IEEE Tranasactions on Evolutionary Computation.*2014;**18:**20-35.
2. Saha, S., Bandyopadhyay, S.,. A Symmentry based Multiobjecive Clustering Technique for Automatic Evolutionary Computation". *Pattern Recognition*. Elsevier. 2010;738-751.
3. Dirk Sudholt. A New Method for Lower Bounds on the Running Time of Evolutionary Algorithms. *IEEE Transactions on Evolutionary Computation*.2013; 418-435.
4. Anirban Mukhopadhyay, Ujjwa Maulik, Sanghamitra Bandyopadhyay. An Interactive Approach to Multiobjective Clustering of Gene Expression Patterns. *IEEE Transactions on Biomedical Engineering.*2013; 35-41.
5. Batista, L., Campelo, F., Guimarães, F., J. Ramírez, J. Pareto cone e-dominance: improving convergence and diversity in multiobjective evolutionary algorithms. *Evolutionary Multi-Criterion Optimization*.2011; 76–90.
6. Xinjie Yu, Mitsuo Gen. Introduction to Evolutionary Algorithms. *Springer-Verlag.* London. 2010.
7. Saber, M., Elsayed, Ruhul, A., Daryl, L., Essam. A New Genetic Algorithm for Solving Optimization Problems. *Engineering Applications of Artificial Intelligence*. Elsvier.2014; 57-69.
8. Dipankar Dutta, Paramartha Dutta, Jaya Sil. Simultaneous Continuous Feature Selection and K Clustering by Multi Objective Genetic Algorithm.*3rd IEEE International Advance Computing Confernece(IAAC).*DOI.978-1-4673-4529-3/12. 2013; 937-942.
9. Anusha, M., Sathiaseelan, J., G., R. An Enhanced K-Means Genetic Algorithm for Optimal Clustering.*2nd IEEE International Conference on Computational Intelligence and Computing Research (ICCIC).*DOI.978-1-4799-3974-9/14.2014; 580-584.
10. Kung-Jeng Wang, Kun-Hang Chen, Angelia, M., Adrian. An Improved Artificial Immune Recognition System with the Opposite Sign Test for Feature Selection. 2014; 1-47.
11. Girish Chandrashekar, Ferat Sahin. A Survey on Feature Selection Methods. 2014; 16-28.
12. Chih-Fong Tsai, William Eberle, Chi-Yuan Chu. Genetic Algorithms in Feature and Instance Selection. *Knowledge-Based Systems.*Elsvier.2013; 240-247.
13. Xu Xia, Jian Zhuang, Dehong Yu. Novel Soft Subspace Clustering with Multi-objective Evolutionary Approach for High-Dimensional Data. *Pattern Recognition.*Elsvier.2013; 2562-2575.
14. Nibaran Das, Ram Sarkar, Subhadip Basu, Mahantaps Kundu, Mita Nasipipuri, Dipak Kumar Basu. A Genetic Algorithm based Region Sampling for Selection of Local Features in Handwritten Digit Recognition Application. *Applied Soft Computing*. Elsevier2012; 1592-1606
15. Pedrycz, W., Syed Ahamad, S., S. Evoutionary Feature Selection via Structure Retention. *Expert System with Applications*. Elsevier. 2012; 11801-11807.
16. Emel Kizilkaya Aydogan, Ismail Karaoglan, Panos, M., Pardalos. hGA: Hybrid genetic Algorithm in fuzzy rule-based classification Systems for High-Dimensional Problems. *Applied Soft Computing.* Elsevier. 2012; 800-806.
17. Ratta, G., A., Vega, J., Muraari, A. Improved Feature Selection on Genetic Algorithms for Real Time Disruption Prediction on JET. Fusion Engineering and Design. Elsevier. 2012; 1670-1678.