Asian Journal of Computer Science and Technology
ISSN: 2249-0701 Vol.7 No.3, 2018, pp. 68-71
© The Research Publication, www.trp.org.in

Multi-Objective Optimization to Identify High Quality Clusters with Close Referential Point using Evolutionary Clustering Techniques

M. Anusha
Department of Computer Science, National College, Trichy, Tamilnadu, India
E-Mail: anusha260505@gmail.com

Abstract - **Most of the real-world optimization problems have multiple objectives to deal with. Satisfying one objective at a time may lead to the huge deviation in other. This paper uses criterion knowledge ranking algorithm solving multi-objective optimization problems. The aim of this research paper is to solve a multi-objective optimization algorithm with close reference point learning method to identify high quality data clusters. A Simple crossover measure is used to quantify the diversity of the whole set, by considering all patterns as a complete entity. In this paper, the task of identifying high quality data clusters using close reference points is proposed to solve multi-objective optimization problem using evolutionary clustering techniques. The proposed algorithm finds the closest feature from the selected features of the data sets that also minimizes the cost while maintains the quality of the solution by producing better convergence. The resultant clusters were analysed and validated using cluster validity indexes. The proposed algorithm is tested with several UCI real-life data sets. The experimental results substantiates that the algorithm is efficient and robust.**

Keywords: **Multi-Objective Optimization, Reference Point Learning, Evolutionary Clustering, Feature Selection, High Quality Data Clusters**

I. INTRODUCTION

Now-a-days, the researchers are directing their focus to solve multi-objective optimization (MOP) problem. Since the solution to the problem is highly signified in the fields of engineering, computer science, information technology and so on [1]. Most of the multi-objective optimization problems have multiple objectives that conflicts each other. Hence, the objectives have to be simultaneously optimized. There are three types are MOPs namely, weighted function, lexicographical approach and Pareto approach. The weighted function is the process of transforming the multi-objective problem into single-objective problem by using certain weighted measure. Lexicographical approach uses ranking methods whereas Pareto approaches consists of several non-dominated solutions [2].

In-order-to solve MOP, there is a need set of optimal solutions, called Pareto-front solutions. These solutions are obtained using Genetic Algorithms (GA) which is subjected to the shape or continuity of the Pareto-fronts. Therefore Pareto-front gained more attention in the field of data mining. Pareto-fronts could be obtained through genetic algorithm. In-order-to mine MOP, there is a need to formulate certain rules such as defining fitness function, adopting accurate genetic operators and so on [3].

The high quality pattern mining is the process of discovering p patterns with the largest utility value from transaction database which has attracted a lot of research work in the field of data mining [4, 5, and 6]. In the task of utility mining, each item is associated with a utility and its quantity in different transactions. The importance of a pattern can be measured by its utility in terms of value or other information stated by users [7]. Most of the existing works for mining high quality data clusters concentres on improving the efficiency of the mining algorithms and the mined patterns that are considered separately during the mining process [8]. Thus, the identified clusters may be very similar and lack diversity. Decision maker system should consider both object and its value as pattern to improve overall satisfaction of users [9, 10]. Hence by considering both value and diversity, this paper proposes to identify high quality data clusters for better convergence result. In this problem, the chosen high quality data clusters are considered as a complete population, gpr. A simple measure of convergence called evolutionary operator is used to quantify the diversity. By considering both object and its value, high quality clusters are obtained even in diversified state. In other words, a higher diverse of P will lead to poor value, whereas a lower diverse of P often leads to higher value. Due to the conflicting property between object and value, reference point based multi-objective evolutionary approach is then proposed to obtain high quality data clusters using close reference points. The rest of the paper is listed as follows. In Section II, a massive study of literature is reviewed. In Section III, the proposed algorithm is discussed in detail. Section IV shows the experimental results obtained from the study. Finally, conclusion and possible research issues are presented in Section V.

II. LITERATURE SURVEY

An extensive survey on MOGA for clustering using reference point based is discussed in this Section. Rui Wang *et al.* [11] applied certain genetic variation for environment selection. HV measures to find the volume of the space dominated by threshold along with one reference point. An island model is introduced as a generalized operation with two and clustering is obtained using IGD metric. This algorithm fails in the combinatorial approach with high diverge cluster solution. Alvaro Gracia-Piquer *et al.* [12] derived PESA II with two validity indices: overall cluster deviation and cluster connectedness. Having used the index

Namely Davies–Bouldin index, the Dunn's index and the Silhoutee index for identifying the cluster shape. The initial class of the data set is modeled using adjusted Rand Index. It is pointed out that bloat-control techniques increases the accuracy with rise in run time. Cluster-merging is a post-processing stage for the final Pareto set which aims to promote cluster separation without penalizing the compactness for the low density clusters respectively. This method lacks in identifying the objectives in the search space.

Hu Xia *et al.* [13] adopted NSGA II for their proposed work with the minimum of two objectives, Jin for the within-cluster dispersion and Jadd for evaluating cluster result associated with Projection similarity-gap-statistic index for obtaining best solution. It is inferred that time complexity is irreverent to the subject space with less cluster compactness. Sujoy Chatterjee *et al.* [14] implemented NSGA II and exemplify the algorithm with the maximized number of cluster similarity with adjusted rand index to the most similar cluster. The algorithm fails in cluster convergence.

Considering MOE/A as basis, Shi-Zheng *et al.* [15] proposed an ensemble clustering which concentrates on neighborhood size. This method uses certain learning process which decomposes the functional value into scalar optimization sub-problem by accompanying IGD metric as its performance metric. There is need to improve in cluster variance. David *et al.* [16] postulated dynamic recommender system based on evolutionary clustering which improves the prediction accuracy also alleviating the scalability of the temporal dimensions. Nevertheless, computational efficiency has to be improved. Hemant *et al.* [17] has developed Pareto corner search evolutionary algorithm which fetches corners of the Pareto front to identify relevant objectives. It is remarked that there was an inadequate use of selection operator that in turn fails in accuracy. Anusha *et al.* [18, 19] focused on centroid based multi-objective clustering that lacks to reduce the number of clusters. Various feature selection and neighbourhood learning techniques are explained in [20, 21]

Zihayat *et al.* [22] and a proposed an efficient algorithm THUDS for mining top-k high utility patterns over data streams. Ryang *et al.* [23] proposed the REPT algorithm with four strategies for efficient top-k high utility pattern mining. Yang *et al.* [24] proposed to mine diversified temporal subgraph patterns, where the coverage set was used for measuring the diversity of a set of temporal subgraphs.Tseng *et al.* [25] proposed two efficient algorithms named TKU and TKO for mining top-k high utility patterns. Leading to that the recommended top-k patterns are very similar and lack diversity. Kannimuthu *et al.* [26] presented the use GA-based algorithm with the ranked mutation to mine high utility patterns .However; it is time consuming to set the appropriate chromosomes at the first stage in their approach. Furthermore, other different evolutionary computation based methods such as particle swarm optimization (PSO) and ant colony system (ACS)) were used for mining high utility patterns effectively and efficiently [27, 28]. The above algorithms were reviewed based on the reference point and cluster closeness is considered not the quality which motivates to develop the task of identifying high quality data clusters using close reference points is proposed to multi-objective optimization problem using evolutionary clustering techniques for better convergence of populations which is extension of previous work.

III. PROPOSED ALGORITHM

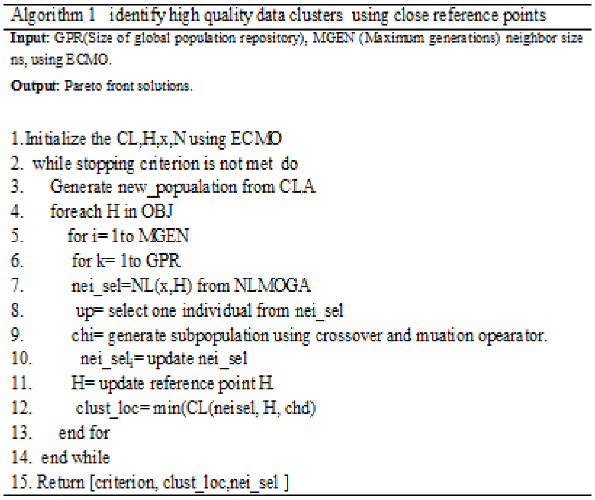
This section proposes a high quality data clusters with the help of close reference points.

A. The Objectives of QP-ECMO

The clusters are selected in such ay that is close to selected reference points. A measure of coverage is introduced to quantify the diversity of high quality patterns. Then, the task can be formulated as a 2-objective optimization problem since that the two measures object and value are conflicting. The population is selected using ECMO algorithm which assures that the selected subpopulation is good and convergence is maintained using evolutionary operators.

B. Mining High Quality Patterns using Reference Points

The proposed algorithm QP-ECMO is implemented by using criterion knowledge ranking evolutionary algorithm (ECMO) [29] to guarantee a good trade-off between the convergence and diversity of populations during evolution. Let *gpr* be the population size. The proposed reference point based representation is given in Algorithm1 which presents, the population initialization, population evolution and population selection. In the first step, the proposed initialization strategy is suggested to obtain the initial population using NLMOGA[30].



```
Algorithm 1   identify high quality data clusters  using close reference points
Input: GPR(Size of global population repository), MGEN (Maximum generations) neighbor size
ns, using ECMO.
Output: Pareto front solutions.

Initialize the CL,H,x,N using ECMO
while stopping criterion is not met  do
    Generate new_popualation from CLA
    foreach H in OBJ
      for i= 1to MGEN
        for k= 1to GPR
        nei_sel=NL(x,H) from NLMOGA
         up= select one individual from nei_sel
        chi= generate subpopulation using crossover and muation opearator.
          nei_sel$_i$= update nei_sel
        H= update reference point H
         clust_loc= min(CL(neisel, H, chd)
     end for
end while
Return [criterion, clust_loc,nei_sel ]
```

Then, the reference point H is obtained by computing the best values of object and its coverage found in the initial population. The Euclidean distances between any individual rand in current population pop calculated. Then, the set of p individuals from gpr with the nearest Euclidean distance to H are chosen as the neighbours of rand. In the second step, for each individual subpop in gpr (1 ≤ i ≤ popsize), one individual rand is randomly selected from updated poplation up . The two individuals up and rand are used to generate the child chi by using the proposed crossover and mutation operators. If the ranking value of chi is better than an individual rand in gpr, then replace rand with chi and update reference point H. In the last step, the algorithm terminates when the maximum number of generations reaches, and the set of non-dominated solutions from the final population is obtained by fast non-dominated sorting strategy.

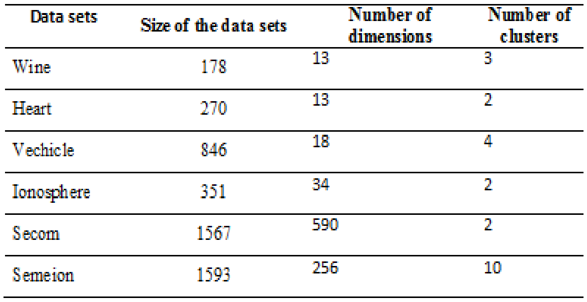
IV. EXPERIMENTAL STUDIES

To evaluate the performance and efficacy of the proposed algorithm QP-ECMO, an unsupervised genetic algorithm is discussed in this section.

A. Data Set and Experimental Setting

The proposed algorithm is tested ample number of microarray datasets. This does not mean that the algorithm works only with microarray datasets, the framework can work well for general clustering problem. Seven real-life dataset are taken from UCI data repository. Table I contains the information about the datasets for the analysis. The algorithm was implemented in 7.6 and executed using Pentium with 2.99 GHZ CPU and 2 GB RAM. The operating system Microsoft Windows XP.

TABLE I INFORMATION ABOUT DATSETS



Data sets	Size of the data sets	Number of dimensions	Number of clusters
Wine	178	13	3
Heart	270	13	2
Vechicle	846	18	4
Ionosphere	351	34	2
Secom	1567	590	2
Semeion	1593	256	10

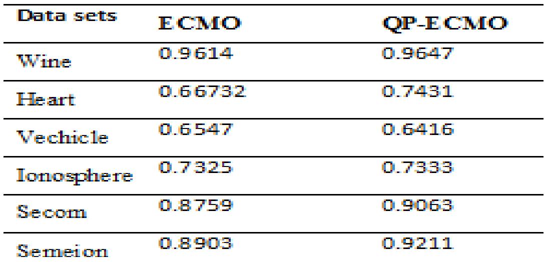
A. Parameter Setting

The number of clusters parameter is fixed for the particular data sets. For the proposed algorithm, The evolution operator is the same for every algorithm, which is the combination of simulated binary crossover (SBX) and the polynomial mutation, with crossover distribution index *cdi* = 20, crossover probability *pc* = 0.95, mutation distribution index *mdi* = 20 and mutation probability *pm* = 1/nei-sel. The performance is measured using rand index (RI) metric which measures both diversity and proximity of an obtained population.

B. Testing Datasets and Performance Metrics

Table II shows the performance metric values obtained by rand index for the seven data sets respectively. It is proved from the Table II, the proposed algorithm QP-ECMO to identify high quality patterns using evolutionary clustering techniques for multi-objective optimization is performing well.

TABLE II EVALUTION OF PROPOSED PERFORMANCE METRIX



Data sets	ECMO	QP-ECMO
Wine	0.9614	0.9647
Heart	0.66732	0.7431
Vechicle	0.6547	0.6416
Ionosphere	0.7325	0.7333
Secom	0.8759	0.9063
Semeion	0.8903	0.9211

V. RESULTS AND DISCUSSION

Let n be the number of transactions in global population repository and i be the number of distinct items in *mgen*, *k* be the number of clusters, *mgen* be the maximum number of generations, *pop* be the size of population and *nei-sel* be the average length of data points in the cluster. The major computational complexity of QP-ECMO is taken for the calculation of objective function values of individuals in the step of population (pop ∗ k ∗ mgen∗ nei-size). The efficiency of QP-ECMO for identifying high quality data clusters is given in Table II. The running time of QP_ECMO on the seven datasets, averaging on 28 runs Hence, the computational cost complexity is O| (GEN*nei-sel)/GPR|. Therefore, the algorithm generates minimum cost for certain data sets.

VI. CONCLUSION

In this paper, a new evolutionary clustering algorithm called QP-ECMO was proposed to address the clustering problem of identifying high quality data clusters using close reference point. Also the issue have been considered as a promising multi-objective clustering problem too. Although, few proposals have been presented in this area, this article differs in identifying high quality data clusters using close reference points. The proposed algorithm produces good convergence with the help of evolutionary operators. The effectiveness and efficiency of the algorithm are verified on seven real world data sets. Experimental results show that it outperforms the previous algorithm ECMO by achieving good convergence to produce high quality data cluster using close reference point. The future direction could be to identify overlapping clusters while generating quality data clusters with different distance metric.

REFERENCES

[1] H.R.Cheshmehgaz, H.Haron, and A.Sharifi, “The review of multiple evolutionary searchs and multi-objecive evolutionary algorithm”, *Artificial Intelligence*, pp. 1-33, 2013.

[2] O.Schutze, M.Laumanns, C.A.C.Collo and E.G.Talbi, “Compting gap free pareto front approximations with stochastic search algorithms”, *Evolu.Compt*., Vol. 18,No. 1,pp. 65-96, 2010.

[3] Sk.M.Islam, S.Das, S.Ghosh, S.Roy and P.N.Suganthan, “An adaptive differential evolution algorithm with novel mutation and

crossover strategies for global numerical optimization", *IEEE. Trans. SMC.*, Vol. 12, No. 2, pp. 282-500, 2012.

[4] Wu, Le, Liu, Qi, Chen, Enhong, Yuan, Nicholas Jing, Guo, Guangming, Xie, Xing, "Relevance meets coverage: A unified framework to generate diversified recommendations", *ACM Trans. Intell. Syst. Technol*. Vol.7, No.3, pp. 39.1-39.30, 2016

[5] Yin, Junfu, Zheng, Zhigang, Cao, Longbing, Song, Yin, Wei, Wei, "Efficiently mining top-k high utility sequential patterns", In: *IEEE International Conference on DataMining*, pp. 1259–1264. 2013.

[6] Zihayat, Morteza, An, Aijun, "Mining top-k high utility patterns over data streams",*Inf. Sci*, Vol.285,No.20,pp.138–161, 2014.

[7] Ryang, Heungmo, Yun, Unil, "Top-k high utility pattern mining with effective threshold raising strategies" *Knowl.-Based Syst*, Vol.76, pp.109–126. 2015.

[8] Tseng, S. Vincent Wu, Chengwei, Fournierviger, Philippe, Yu, Philip S., "Efficient algorithms for mining top-k high utility itemsets", *IEEE Trans. Knowl. Data Eng*, Vol. 28, No. 1, pp. 54–67. 2016.

[9] Hammar, Mikael, Karlsson, Robin, Nilsson, Bengt J., "Using maximum coverage to optimize recommendation systems in e-commerce.", *Proceedings of the 7th ACM Conference on Recommender Systems,* pp. 265–272, 2013.

[10] Lucas, Tarcsio, Silva, Tlio C. P.B., Vimieiro, Renato, Ludermir, Teresa B., "A new evolutionary algorithm for mining top-k discriminative patterns in high dimensional data." *Appl. Soft Comput*, Vol. 59, pp. 487-499. 2017.

[11] Rui Wang, Peter J.Fleming, and Robin C. Purshouse "General framework for localised multi-objective evolutionary algorithms", *Information Sciences*, Elsevier, pp. 29-53, 2014.

[12] Alvaro Gracia-Piquer, Albert Fornells, Jaume Bacardit, Albert Orriols and Elisabet Golobardes,"Large-Scale Experimental Evaluation of Cluster Representations for Multiobjective Evolutionary Clustering", *IEEE Transcations on Evolutionary Computation*, pp. 36-53, 2014.

[13] Hu Xia, Jian Zhuang, Dehong Yu, "Novel soft subspace clustering with multi-objective evolutionary approach for high- dimensional data", *Pattern Recognition, Elsevi*er, pp.2562-2575, 2013.

[14] Sujoy Chatterjee and Anirban Mukhopadhyay, "Clustering Ensemble: A Multiobjective Genetic Algorithm based Approach", *Procedia Technology, Elsevier*, pp. 443-449, 2013.

[15] Shi-Zheng Zhao, Ponnuthurai Nagaratnam Suganthan and Qingfu, "Decomposition–Based Multiobjective Evolutionary Algorithm with an Ensemble of Neighborhood Sizes", *IEEE Transactions on Evolutionary Computation*, pp. 442-446, 2012.

[16] Gema Bello-Orgaz, and David Camacho, "Evolutionary clustering algorithm for community detection using graph-based information", *IEEE Cong.Evolu.Compt.*, pp. 930-937. 2014.

[17] Hemant Kumar Singh, Amitay Isaacs,and Tapabrata Ray, "A pareto corner search evolutionary algorithm and dimensionality reduction in many-objective optimization problem", *IEEE Trans.Evolu.Compt.*, Vol. 15, No. 4, pp. 539-556, 2011.

[18] M.Anusha and J.G.R.Sathiaseelan, "An Improved K-Means Genetic Algorithm for Multi-objective Optimization", *International Journal of Applied Engineering Research*,pp. 228-231, 2015.

[19] M.Anusha and J.G.R.Sathiaseelan, "An Empirical Study on Multi-Objective Genetic Algorithms using Clustering Techniques", *International Journal of Advanced Intelligence Paradigms*. Vol. 8, No. 3, pp. 343-354, UK, 2016.

[20] M.Anusha and J.G.R .Sathiaseelan, "Feature Selection using K-Means Genetic Algorithm for Multi-objective Optimization", *Procedia Computer Science*, Vol. 57, pp. 1074-1080. Elsevier B.V., Netherland, 2015

[21] M.Anusha and J.G.R.Sathiaseelan, "An Enhanced K-means Genetic Algorithms for Optimal Clustering", *IEEE International Conference on Computational Intelligence and Computing Research (ICCIC),* pp.580-584, 2014.

[22] Zihayat, Morteza, and Aijun, "Mining top-k high utility patterns over data streams", *Inf. Sci*.Vol. 285, No.20, pp. 138–161, 2014.

[23] Ryang, Heungmo, Yun, Unil, "Top-k high utility pattern mining with effective threshold raising strategies", *Knowl.-Based Syst.* Vol.76, pp.109–126. 2015.

[24] Yang, Yi, Yan, Da, Wu, Huanhuan, Cheng, James, Zhou, Shuigeng, Lui, JohnC.S., "Diversified temporal subgraph pattern mining", *Proceedings of ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 1965–1974. 2016.

[25] Tseng, Vincent S., Wu, Cheng-Wei, Shie, Bai-En, Yu, Philip S., "UP-growth: an efficient algorithm for high utility item set mining", *Proceedings of ACM SIGKDD Conference on Knowledge Discovery and Data Mining. ACM*, pp. 253–262, 2010.

[26] Kannimuthu, S., Premalatha, K., "Discovery of high utility itemsets using genetic algorithm with ranked mutation" *Appl. Artif. Intell*, Vol. 28, No.4, pp. 337–359, 2014.

[27] Lin, Jerry Chun-Wei, Yang, Lu, Fournier-Viger, Philippe, Hong, Tzung-Pei, Voznak, Miroslav, "A binary PSO approach to mine high-utility itemsets", *Soft Comput*. Vol. 21, pp. 5103–5121. 2017.

[28] Wu, Jimmy Mingtai, Zhan, Justin, Lin, Jerry Chunwei, "An ACO-based approach to mine high-utility itemsets", *Knowl.-Based Syst*, 2017. Vol.116, pp.102–113. 2017.

[29] M.Anusha and J.G.R. Sathiaseelan, "Evolutionary Clustering Algorithm using Criterion-Knowledge-Ranking for Multi-objective Optimization", *Wireless Personal Communication*, Springer, Vol.94, pp.2009-2030, Springer, USA. 2017

[30] M.Anusha and J.G.R. Sathiaseelan, "An Improved K-Means Genetic Algorithm for Multi-objective Optimization", *International Journal of Applied Engineering Research*, pp. 228-231, 2015.