Multi-Objective Optimization Algorithm to the Analyses of Diabetes Disease Diagnosis

M. Anusha
Dept. of Computer Science
Bishop heber college
Trichy, India. 620017

Dr. J.G.R. Sathiaseelan
Dept. of Computer Science
Bishop heber college
Trichy, India.620017

***Abstract*—There is huge amount of data available in health industry which is found difficult in handing, hence mining of data is necessary to innovate the hidden patterns and their relevant features. Recently, many researchers have devoted to the study of using data mining on disease diagnosis. Mining bio-medical data is one of the predominant research area where evolutionary algorithms and clustering techniques are emphasized in diabetes disease diagnosis. Therefore, this research focuses on application of evolution clustering multi-objective optimization algorithm (ECMO) to analyze the data of patients suffering from diabetes disease. The main objective of this work is to maximize the prediction accuracy of cluster and computation efficiency along with minimum cost for data clustering. The experimental results prove that this application has attained maximum accuracy for dataset of Pima Indians Diabetes from UCI repository. In this way, by analyzing the three objectives, ECMO could achieve best Pareto fronts.**

Keywords—Clustering; Genetic Algorithm; Multi-objective Optimization; ECMO; Diabetes Disease

I. INTRODUCTION

Among numerous diseases, health department of India have identified that Diabetes disease lists top for cause of death domestically. In the recent years, it is inferred that this problem is growing at an alarm rate with massive patient data [1]. In this view, this work adopts Evolutionary Clustering Multi-objective Optimization Algorithm (ECMO) which was the extended work of NL-MOGA for analyzing diabetes disease datasets [2]. There are some global optimization tools such as genetic algorithms uses validity measures for evaluating clustering accuracy [3]. However, as no single validity measures works equally well for different datasets which could simultaneously produce high clustering accuracy. Some recent studies have posed the problem of data clustering as a multi-objective optimization problem in which several cluster validity measures are optimized concurrently to obtain the tradeoff clustering solutions.

Depending on the dataset properties and its inherent clustering structure, different cluster validity measures perform differently [4]. Therefore, it is important to find the best validity measures that could be instantaneously attain good clustering results. In order to evaluate the quality of the clustering, external measures like Jaccard-index, Minkowski-index, Rand-index, and so on can be utilized to optimize the multi-objective problem [5]. This measure are used to identify the intra-cluster similarity or compactness and the inter-cluster separation. In this paper, the cluster compactness and the separation is evaluated using Rand- index. This index measures both cohesion and separation of clusters using distance measures between the points in the closest cluster to the points in the same cluster [6]. The Rand-index for the point xi is calculated as

$$R(T,C) = \frac{a+d}{a+b+c+d} \quad (1)$$

Whereas, T is the true cluster of the selected dataset for C the clustering result returned by some algorithm. The points a, b, c, and d are the objects belonging to T and C. The value close to +1 indicates a good clustering. Hence, the best cluster accuracy can be calculated using this index. Thus the inaccuracy could be the values nearing -1[7]. However, in clinical diagnosis, the inaccuracy could be in diagnosing false positivity and negativity.

The false positive like inaccurate-positive depicts the percentage of patients diagnosed to have no disease while in reality they have. Inaccurate-negative implies that the ratio of patients being diagnosed with disease but is diagnosed contrarily. In general, false negative results could cause greater impact than false positive results for both the doctors and the patients. At this juncture, the analysis of medical disease needs more concentration on the state of lower cost or false negative than the state of high cost or false positive.

Therefore by applying ECMO algorithm which uses data mining technology along with genetic algorithm that would help in analyzing the disease to produce high accuracy results by optimizing the low cost and high cost values. In this light, the accuracy and cost are the conflicted objectives. Hence, the optimum results could be achieved by setting minimum acceptable accuracy rate. On the premium that all the conditions were attained, the higher accuracy and low cost values could result better. The optimum values could be drawn from Pareto fronts. The rest of the paper is organized as follows. In Section II, a brief review of some past studied are presented. Then in Section III, methodology of ECMO is discussed in detail. Section IVshows the experimental results obtained from the study. Finally, conclusion and possible research issues are presented in Section V.

II. LITERATURE REVIEW

Sriparna et al. [8] proposed a multi-objective clustering technique to partition the data into appropriate clusters. This work aims to find total compactness of the partitioned clusters, symmetry of the clusters and the connectedness of the clusters. The algorithm uses Silhouette-index to measure the validity of

the clusters. Hector et al. [9] presented a technique to identify the main folds in the large datasets. Author summarized the original search space with Map-reduce architecture to identify the voronoi regions. Guang et al. [10] depicted generate-first-choose next method using upper bounds, lower bounds and inequality constraint engineering problem based on surrogate models. The algorithm failed to adopt weighted sum approach.

Lei et al. [11] devised clustering-ranking algorithm using a series of reference lines as cluster centroid. The solutions are ranked according. Anibran et al. [12] defined an interactive genetic algorithm based multi-objective approach that could simultaneously found clustering solution by evaluating the validity measures. The algorithm reduces fatigue of the decision maker by generating only important solutions from the current population. A massive on clustering based multi-objective genetic algorithm is presented in [13] and the author extended research by depicting an enhanced K-means Genetic algorithm for optimal clustering. The author overcomes the drawback of local optima with suitable dataset and also the algorithm fails in computational time. It is inferred that the algorithm produced more than the 90% accuracy for real life dataset. The author also adopted a neighborhood learning strategy for optimizing multi objective problems. This algorithm used k means Genetic algorithm to find the compactness of the clusters. It is noted that the algorithm could produce minimum index value for the maximum datasets. However, there is a need for proper feature selection for better, more optimal solution [14, 15]. Ruby et al. [16] suggested two methods for ranking of MOPs. This ranking methods were used to prune large data-sets of solution to small subset of good solution. Edward et al. [17] presented an approach by extracting the knowledge of conflicting interests like traceability and transparency to obtain the group of consensus data. Min Han et al. [18] considered mutual information based feature selection to enhance the searching capability of the data. Partha Pratim et al. [19] proposed high dimensional feature selection technique to preserve sample similarity using shared neighbor distance technique to reduce the outliers with a minimum computational complexity.

III. Methodology

This section address the issues specified in Section II by applying evolutionary clustering algorithm (ECMO) for MOPs. Primarily, ECMO generates uniform set of objects as the population. Then, the population is treated with three main procedures until the termination condition is satisfied. The three major operations are criterion learning algorithm (CLA), knowledge acquisition algorithm (KAA) and optimal cluster-ranking algorithm (RA). The ultimate goal of CLA is to perform global search based on the discovered criteria and then the knowledge is acquired through constant learning to dominance. While RA refine the process by grouping most relevant data with the help of ranking strategy.

A. Evolutionary Clustering Algorithm for Multi-objective Optimization

This research inherits ECMO which handles data by adopting criterion learning algorithm. The criterion for the particular objective was designed based on cluster location. The neighborhood data such as closest neighbor, farthest neighbor and indirect neighbor were identified using knowledge acquisition algorithm. Hence, based on the dominance of individuals the data can be grouped and ranked using best knowledge ranking algorithm. The optimal Pareto fronts was achieved using balancing Pareto front algorithm that was capable of finding the best features the particular data set. Therefore, the fitness function for diabetes disease diagnosis using ECMO could be maximizing the cluster accuracy with minimum number of false negatives and false positives. It can be represented as follows:

$$f_1(x) = Max \sum_{i=1}^{n} |C(_{origi}x_i) + C(_{pred}x_i)| \quad (2)$$

$$f_2(x) = Min \sum_{i=1}^{n} (|Inaccu_{negative}|) \quad (3)$$

$$f_3(x) = Min \sum_{i=1}^{n} (|Inaccu_{positive}|) \quad (4)$$

Hence, by adopting the rules of knowledge acquisition algorithm true negatives and true positives objects can be identified. Maximum cluster accuracy could be achieved through best knowledge ranking algorithm of ECMO.

IV. Experimental Studies

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

A. Data Set and Experimental Setting

The algorithm is tested Pima Indian Diabetes microarray datasets which are taken from UCI repository [20]. There are 768 records, out of which 268 cases are with diabetes disease and 500 cases are without diabetes with 376 records contain missing values. Pima Indian Diabetes microarray datasets contains 8 attributes with on class attribute. Table I contains the information about the dataset for the analysis. The algorithm were 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 Attributs of Diabetes Disease

No	Attributes	Domain
1	Age	continuous
2	No. of times of pregancy	0,1,2
3	Diastolic Blood pressure	continuous
4	Plasma glucose concentaion	continuous
5	Triceps skin fold thickness	0,1
6	2-hrs serum insulin	continuous
7	Body mass index	0,1
8	Diabetes pedigree function	continuous
9	Class	Healthy/Sick

B. Testing Datasets and Performance Metrics

The experiment on the dataset was conducted on 90% of training dataset with 10% of test data. Testing has undergone 20 independent runs. The foremost aim of cluster validity indexes is to validate clustering solution. This index is useful in comparing the performance of the cluster. We adopted rand index (RI) to compare the performance of the algorithm with the selected diabetes datasets. The cluster accuracy, inaccurate positive and inaccurate negative for predicting diabetes disease is shown below:

$$Rand_{Clus_{Accu}} = \frac{sick+healty}{sick+indirt_{nei}+fart_{nei}+healthy} \quad (5)$$

$$Inaccu_{negative} = \frac{sick}{\sum_{i=1}^{n} C(sick-indirt_{nei}) + \sum_{i=1}^{c} C(healthy-far_{nei})} \quad (6)$$

$$Inaccu_{positive} = \frac{healthy}{\sum_{i=1}^{n} C(healthy-fart_{nei}) + \sum_{i=1}^{c} C(sick-indirt_{nei})} \quad (7)$$

After eliminating the missing values using mutation operator, the testing of data starts with phase training samples.

1) TEST: 90% of testing dataset (353 cases) and 10% training dataset (39 cases) of 392dataset.

In this test phase, 353 cases training sets with 39 cases of testing samples are considered. During each run, ECMO select different features from the original attributes and the clustering accuracy is recorded. The experiment was repeated 20 times and the results are recorded in Table II.

TABLE II. AVERAGE CLUSTER ACCURACY FOR 20 RUNS

Test Run	No. of Attributes Selected	Rand-Index (%)
1	5	97.65
2	7	96.34
3	8	98.21
4	4	98.49
5	6	99.92
6	8	97.01
7	7	97.57
8	7	98.53
9	6	99.01
10	7	98.76
11	4	96.32
12	5	99.67
13	6	99.46
14	8	98.99
15	8	98.52
16	7	99.21
17	4	99.05
18	3	98.47
19	3	99.38
20	6	98.25

It is inferred from the result that the average closer cluster accuracy is determined using rand index metric. The average clustering accuracy is 98.48%. The results of Pareto fronts was presented in Fig.1. shows the best cluster accuracies produced by the selected objectives. Blue color implies the healthy objects whereas pink and yellow color indicated inaccurate negative and inaccurate positive respectively. The evaluation metrics obtained by ECMO algorithm is recorded in Table III. The Fig. 2 Shows the best Pareto fronts obtained by the selected class variables for the single run. The selected from the Pareto fronts were mostly in the knee regions of the Pareto fronts.

It is noted that cluster prediction the algorithm could able to produce accurate cluster classification with low inaccurate positive and negative results. Table IV represents the impact of ECMO on inaccurate negative and positive results.

ECMO takes 20 iterations independently on diabetes dataset for its clustering process. It is praiseworthy that ECMO could form cluster along with good convergence and diversity as shown Fig.1. It is observed from Fig.2. ECMO can produce Pareto optimal solution for the selected objectives.

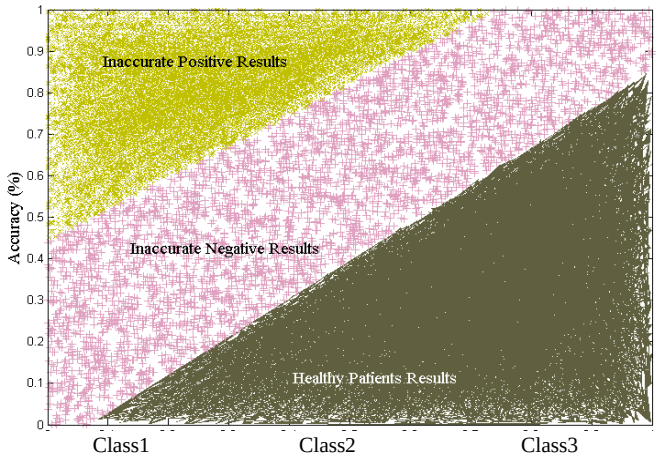

Fig. 1. Cluster Accuracies of the selected objectives

TABLE III. PERFORMANCE MATRICS OBTAINED BY RAND-INDEX

No	Attributes	Domain
1	No. of Attributes	8
2	$Inaccu_{negative}$	83.27%
3	$Inaccu_{positive}$	74.87%
4	$Rand_{Clus_{Accu}}$	98.48%
5	No. of False Negative	5
6	No. of False Positive	10

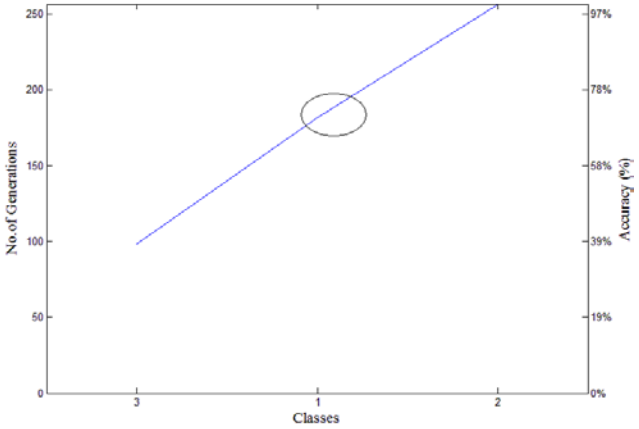

Fig. 2. Cluster Accuracies of the selected objectives

It can be identified from the Table III rand index value of the proposed algorithm is comparatively low than other algorithms except few. When the value of RI is equal to 1, the formation of cluster will be good. Hence, it is certain that ECMO generates better convergence and diversity. Experimental results substantiates that the algorithm ECMO, can identify appropriate features set using criterion and produces better clusters by utilizing the procuring the knowledge from the neighbors. The algorithm adopts neighborhood learning from the previous work and the NLMOGA procedure is extended to figure the closest-neighbor, farthest-neighbor and the indirect neighbor. Based on the outcomes of CLA and KAA, excellent clusters were ranked with more compact and less in diversity. The Table IV reveals that performance of ECMO on healthy, inaccurate positive and inaccurate negative results for diagnosis of diabetes disease.

TABLE IV. COMPARISON OF $Inaccu_{negative}$ AND $Inaccu_{positive}$ USING ECMO

Original \ Prediction		Healthy	Sick
Healthy	Inaccurate Positive	305	48
	Inaccurate Negative	325	29
Sick	Inaccurate Positive	5	34
	Inaccurate Negative	10	29

Hence, it was inferred that the algorithm selected minimum five attributes and the maximum of eight attributes as its feature to process the objective function. It was also noted that the algorithm could able to produce maximum accuracy of 99.92% at the 5th iteration.

Total number of false negative and false positives was noted to very minimum. Therefore, the ECMO produced high cluster accuracy at minimum computation time. Henceforth, it was recorded that the algorithm ECMO produced maximum cluster accuracy for the healthy dataset of disease diabetes by minimizing the inaccurate positive and inaccurate negative results in minimum CPU running time that could reduce the cost substantially.

V. CONCLUSION

This research application on diagnosing diabetes disease using evolutionary clustering multi-objective algorithm which helps in analyzing the datasets found in Pima Indian Diabetes datasets of UCI repository. In this work, the best feature of the dataset was identified using selecting features (CL) of criterion learning algorithm.

The inaccurate positive and inaccurate negative neighbors were identified using knowledge acquisition algorithm. Hence, the algorithm could able to recognize more suitable healthy and sick objects while it possesses the similar dissimilar properties from the selected feature respectively.

ECMO shifts the objects position according to their relative proximity. Hence, the experimental results recorded the optimal solution with good Pareto fronts and high accuracy in healthy clustering. The algorithm could able to produce better cluster accuracy in identifying the inaccurate positive and negative results. Therefore, the reliability by satisfying the considered objectives. Also, algorithm can predict appropriate number of clusters for all the three objectives respectively. Much further work is needed to investigate the utility of having different and more objectives and to test the approach still more extensively, to investigate the utility of having different and more objectives, to hybrid ECMO with multi-objective Particle Swarm Optimization technique for high effectiveness, efficiency, and consistency and to enhance with heterogeneous data.

REFERENCES

[1] Sapna S, Tamilarasi A, and Pravin kumar M, "Implementation of Genetic Algorithm in Predicting Diabetes", International Journal of Computer Science, vol 9, no 3, pp. 234-240, 2012.

[2] Anusha M, and Sathiaseelan, "Evolutionary Clustering Algorithm using Criterion-Knowledge-Ranking for Multi-objective Optimization", unpublished.

[3] Linzhong Liu, Haibo Mu,and Juhua Yang, "Generic constraints handling techniques in constrained multi-criteria optimization and its application.", European J.Operat.Reas.,pp. 1-15, 2015.

[4] Ana B.Ruiz, Mariano Luque, Kaisa Miettinen,and Ruben Saborida, "An interactive evolutionary multiobjective optimization method: Interactive WASF-GA.", Springer Internatinal Publishing, pp. 249-263, 2015.

[5] N.Das, R.Sarkar, S.Basu, M.Kundu, M. Nasipuri, D.K.Basu, "A genetic algorithm based region sampling for sampling for selection of local features in handrwritten digit recognition applications", Appl.Soft.Comp.,vol. 12,no. 5, pp. 1592-1606, 2012.

[6] L. Batista, F. Campelo, F. Guimarães, J. Ramírez, "Pareto cone e-dominance: improving convergence and diversity in multiobjective evolutionary algorithms, in: Evolutionary Multi-Criterion Optimization", pp. 76–90, Springer, 2011.

[7] Maulik U and Bandyopahyay S, "Performance evaluation of some clustering algorithms and validity indices", IEEE Transactions on Pattern Analysis, pp. 1650-1654, 2002.

[8] Sriparna Saha, and Sanghamitra Bandyopadhyay, "A generalized automatic clustering algorithm in multiobjective framework", Applied Soft Computing, vol 13, pp.89-108, 2013.

[9] Hector D. Menendez, and David Camacho, "MOGCLA: A mult-objective genetic clustering algorithm for large data analysis", GECCO'15 ACM 978-1-4503-3488-4, pp. 1437-1438, 2015.

[10] Gung yang, Tao Xu, Xiang Li, Haohua Xiu, and Tianshuang Xu, "An Efficient Hybrid Algorithm for Multi-objective Optimization Problems with Upper and Lower Bounds in Engineering", Mathematical Problems in Engineering, pp. 1-13, 2015.

[11] Lei Cai, Shiru Qu, Yuan Yuan, and Xin Yao, "A clustering-ranking method for multi-objective optimization", Applied Soft Computing, vol 35, pp.681-694, 2015.

[12] Anirban Mukhopadhyay, Ujjwal Maulik,and Sanghamitra Bandyopadhyay, "An interactive approach to multiobjective clustering of gene expression patterns.", IEEE Trans.Biomed.Engg., vol. 60, no. 1, pp. 35-41, 2013.

[13] M.Anusha and J.G.R.Sathiaseelan, (in press), "An Empirical Study on Multi-Objective Genetic Algorithms using Clustering Techniques", International Journal of Advanced Intelligence Paradigms.2015.

[14] M.Anusha and J.G.R.Sathiaseelan, "An Enhanced K-means Genetic Algorithms for Optimal Clustering", IEEE ICCIC, pp.580-584, 2014.

[15] 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.

[16] Ruby L.V.Moritz, Enrico Reich, Maik Schwarz, Matthias Bert,and Martin Middendorf, "Refined ranking relaions for selections in multi-objective mataheuristics.", European J.Operat.Reas., pp. 1-11, 2014.

[17] Edward Abel, Ludmil Mikhailov, and John Keane, "Group aggregation comparisons using multi-objective optimization", Information Sciences, vol 322, pp. 257-275, 2015.

[18] Min Han, and Weijie Ren, "Global mutual information-based feature selection approach using single-objective and multi-objective optimization", http"//dx.doi.org/10.1016/j.neucom.2015.06.016, 2015.

[19] Partha Pretim, and Kundu Sushmita Mitra, "Multi-objective Optimization of Shared Nearest Neighor Simiarity for Feature Selection" Applied Soft Computing, http://dx.doi.org/10.1016/j.asoc.2015.08.042, 2015.

[20] UCI Machine library for source input dataset. https://archive.ics.uci.edu/ml/datasets/PimaIndiansDiabetes.