International Journal of Scientific Research in Multidisciplinary Studies
Volume-5, Issue-7, pp.01-07, July (2019)

Research Paper
E-ISSN: 2454-9312
P-ISSN: 2454-6143

An improved data clustering algorithm using NSGA-II

M.G.Pagale[1*], R.S.Hanchate[2]

[1]Department of Computer, Pimpri Chinchwad College of Engineering & Research, SPPU, Pune, India
[2] Department of Computer, D.Y.Patil College of Engineering, SPPU, Pune, India

Available online at: www.isroset.org

Received: 10/May/2019, Accepted: 24/Jun/2019, Online: 30/Jul/2019

***Abstract*—** In Data clustering, there are various Multiobjective clustering techniques evolved which can automatically partition the data into appropriate no of clusters. For achieving multiple objective functions simultaneously Multiobjective optimization technique is used. Three objective functions such as compactness, connectedness and symmetry of the cluster are optimized simultaneously using NSGA-II. The compactness of the cluster is based on Euclidean distance, a point symmetry based distance used to measure the symmetry of the cluster and Connectedness [1] of the cluster is measured by using relative neighborhood graph concept. Sub cluster are merged appropriately to form variable no of global cluster for objective function evaluation. In this method data is partitioned using k-means clustering algorithm and three objective functions such as compactness, symmetry and connectedness of cluster is optimized by using NSGA-II algorithm. In order to get appropriate no of cluster and accurate partitioning Two-Stage genetic algorithm is applied to these three objective functions.

Keywords—*Euclidean distance,Genetic Algorithm,Multiobjective optimization (MOO),Relative neighborhood grap, Symmetry.*

I. INTRODUCTION

CLUSTERING

Clustering is nothing but allocating the population or data points into a number of groups such that data points in the same groups are more similar to data points in the same group than data points in other groups. In simple words, the aim is to separate groups with similar data and assign them into clusters. For clustering, various multiobjective techniques are evolved, which can automatically partition the data into an appropriate no. of clusters. K-means is very popular data clustering algorithm and is proven to be better for many practical applications[2]. Clustering is a well known unsupervised learning problem; so, in clustering for every problem, it deals with finding a structure in a collection of unlabeled data. Technique follows a simple way to classify a given data set through a certain k number of clusters. The first step is to define k centers, one for each cluster. These cluster centers should be placed in a cunning way because result may vary for different locations. So, place them as much as possible far away from each other. The next step is to take each point belonging to a given data set and assign it to the nearest center. When no point is pending, the first step is finalized and an early group age is done. At this time we need to re-calculate k new centroids of the clusters resulting from the previous step. After we have these k new centroids, a new binding has to be done between the same data set points and the nearest new center. A loop has been generated. As a result of this loop we may observe that the k centers change their location step by step until no more changes are done or in other words centers do not move any more.

MULTI-OBJECTIVE OPTIMIZATION

Multi-objective optimization is well-known as multi-objective programming, vector optimization, multicriteria optimization, multiattribute optimization or Pareto optimization. MOO is an area of multiple criteria decision making that is concerned with mathematical optimization problems involving more than one objective function to be optimized simultaneously. Multi-objective optimization is useful in many fields where optimal decisions need to be taken between two or more conflicting objectives such as science, engineering, economics and logistics[3]. There can be more than three objectives for any practical problem.

For a nontrivial multi-objective optimization problem, no single solution exists that simultaneously optimizes each objective. In that case, the objective functions may be conflicting, and there exists a number of Pareto optimal solutions All Pareto optimal solutions are considered equally good as vectors cannot be ordered completely. The aim is to find set of Pareto optimal solutions and quantify the trade-offs in satisfying the different objectives or finding a single

solution that satisfies the subjective preferences of a human decision maker (DM).

K-MEANS CLUSTERING ALGORITHM

K-mean is unsupervised learning technique that resolves the well-known clustering problem. The method follows a simple and easy way to classify a given data set into an appropriate number of clusters (assume k clusters)[10]. The main idea in clustering is to define k centroid for each cluster. These centroid are placed in a cunning way because of different location gives different result. So, it is better to place them as much as possible far away from each other. In next step take each point belonging to a given data set and assign it to the nearest centroid.When no point is pending first step is fineshed. At this stage we need to again re-calculate k new centroid of the clusters resulting from the previous step. A new Assignment has to be done between the same data set points and the nearest new centroid if we have k new centroid. The k centroid step by step changes their location until no more changes are done. K-Means Algorithm Steps: Let us consider, D = {d1,d2,d3,……..,dn} is a set of data points and C = {c1,c2,…….,cc} is set of centers.

Step 1. Randomly select „C" cluster centers.
Step 2. Calculate the distance of each data point from cluster centers.
Step 3. Assign data point to the cluster whose distance from the cluster center is minimum compare to all the cluster centers.
Step 4. Recalculate new cluster center.

Step 5. Again the distance between new obtained cluster centers and each data point are calculated.
Step 6. If data point was not reassigned then stop, otherwise repeat from step 3.

Rest of the paper is organized as follows, Section I contains the introduction of clustering, multi-objective optimization, k-means clustering. Section II contains the related work of Multiobjective optimization algorithm, Section III contains workflow of proposed system and proposed algorithm, Section IV contains result of Multiobjective functions.

II. RELATED WORK

A multi-objective differential evolution technique was proposed, which uses a variant of the original differential evolution to create the offspring and the best individual is adopted . To implement the selection of the best individual A Pareto-based approach is introduced. For a dominated solution, a set of non-dominated individuals can be identified and the "best" turns out to be any individual randomly picked from this set[4].

A simple evolutionary algorithm called the Pareto Archived Evolution Strategy [5] was proposed[PAES]. In PAES by using mutation one parent generates one offspring. The offspring and parent both are compared. If the offspring dominates the parent, then the offspring is accepted as the next parent and the iteration continues. The offspring is discarded if the parent dominates the offspring, and the new offspring is generated[6]. A comparison set of previously nondominated individuals is used if the offspring and the parent do not dominate each other. An archive of nondominated solutions is considered for maintaining population diversity along the Pareto front. The Archive and a new generated offspring are compared to verify if it dominates any member of the archive. If yes, then the offspring is accepted as a new parent. The dominated solutions are also eliminated from the archive. If any member of the archive is not dominated by the offspring, both parent and offspring are checked with the solution of the archive for their nearness. If the offspring resides in the least crowded region in the parameter space, it is accepted as a parent and a copy is added to the archive [10].

The Strength Pareto Evolutionary Algorithm (SPEA)[6] is described in this paper. At every generation the algorithm maintains an external population by storing all the nondominate solutions obtained so far. The external population is mixed with the current population at each generation. Fitness functions are applied to all nondominated solutions in the mixed population based on the number of solutions they dominate. The Dominated solutions are assigned with the worst fitness of any dominated solution. For ensuring diversity among the nondominated solutions a deterministic clustering technique is used.

A variant of SPEA called SPEA2[7] was proposed. SPEA2 consists of two populations. As in the Initial phase the external population is empty. All nondominated solutions from the current and external population are passed in the next population after the fitness evaluation [8]. The Next population is filled with dominates and individuals from the current and external population if the number of these solutions is less than the population size. A Fitness assignment and a truncation operator are the main differences between SPEA and SPEA 2. A new algorithm for multiobjective optimization is called the Adaptive Pareto Algorithm (APA). APA uses a new technique called as the Adaptive Representation Evolutionary Algorithm (AREA)[9]. This technique allows each solution to be encoded over a different alphabet and the representation of a particular solution is not fixed. Representation is adaptive; it can be changed during the search process as the effect of the mutation operator. APA considers a single population of individuals. Each individual is a unique variation operator and it is selected for mutation. Both the offspring and parent are compared. Survival is guided by the Dominance relation. The offspring enters the new population if the offspring dominates the parent and the parent is removed. An effective and

efficient diversity preserving mechanism is generated by an adaptive representation mechanism and the survival strategy.

The Vector Evaluated Genetic Algorithm (VEGA) is the first genetic algorithm used to approximate the Pareto optimal set by a set of non-dominated solutions; it was implemented by Schaffer [10]. VEGA is an extension of a simple genetic algorithm for multiobjective optimization. Since a number of objectives (M) have to be handled, Schaffer thought of dividing the genetic algorithm population into M equal subpopulations randomly, in each iteration. Based on a different objective function each subpopulation is assigned a fitness[11]. In such a way each M objective function is used to evaluate some members in the population. This algorithm provide solutions, which are good for individual objective functions. VEGA uses the proportional In this way, a crossover between two good solutions, each corresponding to a different objective may find offspring's, which are good compromised solutions between the two objectives. The mutation is applied to each individual as usual.

III. PROPOSED WORK

This work will automatically partition the data into an appropriate number of clusters, for that purpose a new multiobjective (MO) clustering technique is proposed. Fig.1. shows the flow of proposed work. For representing the whole clustering each cluster will be divided into a number of small hyper spherical sub clusters and the centers of all these small sub-clusters are encoded in a string. These local sub clusters are considered independently for assigning points to different clusters. For the purpose of objective function evaluation, these sub-clusters will be merged appropriately to form a variable number of global clusters [11]. Three objective functions, based on the Euclidean distance one reflecting the total compactness of the clustering, the other reflecting the total symmetry of the clusters, and third reflecting the cluster connectedness, are considered here. By using a newly developed simulated annealing based multiobjective optimization method AMOSA[12] these are optimized simultaneously, and also using k-means algorithm in order to detect the appropriate number of clusters as well as the appropriate partitioning. In order to get appropriate no of cluster and accurate partitioning Two-Stage genetic algorithm is applied to these three objective functions. The two-stage selection and mutation operations are implemented to exploit the search capability of the algorithm by changing the probabilities of selection and mutation according to the consistence of the number of clusters in the population [13].

Symmetry: A newly developed point symmetry based distance used to measure the symmetry present in a partitioning. point symmetry based cluster validity index, Sym-index, is used as a measure of the validity of the corresponding partitioning[14].

Connectedness: to measure Connectedness present in a partitioning relative neighborhood graph concept used.
Compactness: The total compactness of the partitioning is based on the Euclidean distance.
Thus the proposed system will be able to detect the appropriate number of clusters and the appropriate partitioning from given data sets having either well partitioned clusters of any shape or symmetrical clusters with or without overlaps.

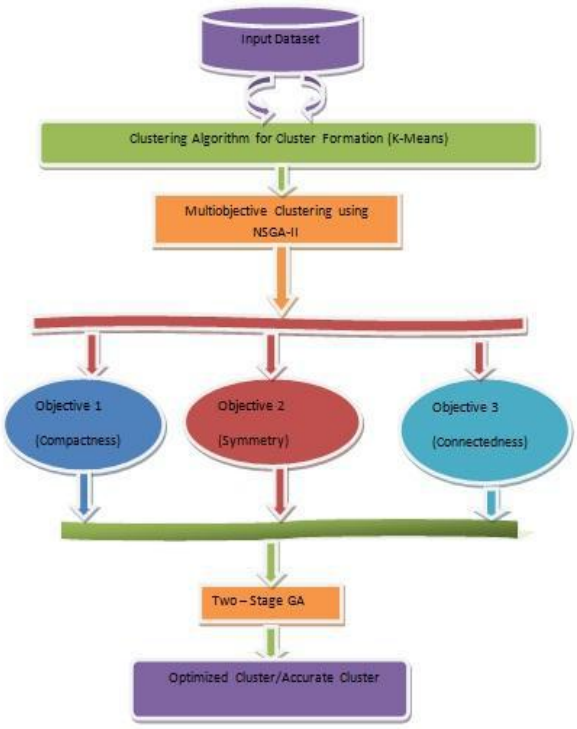

Fig. 1. Workflow of Proposed System

Proposed Algorithm Steps
1. Take input dataset.
2. Data Clustering by using K-means Algorithm
3. Generate Population of clusters
4. Multiobjective Clustering using NSGA-II.
4.1 Set three objective functions.
4.2 Achieve symmetry by point symmetry based Distance.
4.3 Achieve connectedness using the relative Neighbourhood graph concept.
4.4 Achieve Compactness using the Euclidean Distance concept.
5. Apply two stage GA operators
6. Store the optimized or accurate clusters.

IV. RESULTS AND DISCUSSION

A. Dataset Used

For the experimentation, six real life data sets are used from a UCI machine learning repository [15]. These data sets are described in terms of the number of points present, dimensions, and the number of clusters in Table 1.

Table 1. Datasets

Sr. No.	Dataset	classes	features	Size
1	Iris	3	4	150
2	Glass	6	9	214
3	Wine	3	13	178
4	New thyroid	3	5	215
5	Lung cancer	3	57	32
6	Liver Disorder	2	7	345

In order to evaluate the performance of a proposed method, F-measure[15] is used as a performance metric. F-measure values are calculated from the precision and recall values. Precision is the fraction of retrieved instances that are relevant, while recall is the fraction of relevant instances that are retrieved.

Result

In order to detect the appropriate number of clusters and the appropriate partitioning, we implemented the K-means algorithm in a Multiobjective framework using the NSGA II Algorithm. Three objective functions are applied for cluster optimization.

The initial clusters resulted from the K-means algorithm form the population for the NSGAII algorithm. In order to produce the optimized clusters three objective functions are applied one after the other; first, the connected clusters are obtained by using the relative neighborhood graph concept [16], then the compactness of the clusters is calculated by using the Euclidean distance formula and lastly, by using the point symmetry based distance symmetry is calculated. The performance is checked by using a single objective function and multiple objective functions on the same dataset. The F-measure values are as shown in Table 2.

Table 2. Results Of Data Clustering For Single And Multiple Objective Functions.

Multiple-Objective Functions (F-Measure value)				
Input Dataset	**Compactness**	**Connectedness**	**Symmetry**	**Multiple-Objectives**
Cancer	0.969±0.008	0.99±0.002	0.969±0.004	**0.990±0.008**
Iris	0.886±0.004	0.872±0.004	0.873±0.002	**1.0±0.002**
Liver disorder	0.928±0.002	0.892±0.006	0.85±0.006	**1.0± 0.002**
Lung cancer	0.887±0.003	0.912±0.002	0.996±0.004	**0.994±0.003**
New thyroid	0.944±0.001	0.874±0.004	0.883±0.008	**0.998±0.008**
Wine	0.849±0.006	0.842±0.006	0.862±0.002	**0.882±0.004**

B. Result

Table 3 summarizes the F-Measure Obtained for Compactness of cluster from the seven Multiobjective clustering algorithms for the six data sets and the Fig. 2. Shows graph for cluster compactness.
Objective function 1 : Compactness
Performance Evaluation Method : F-Measure[16].

Table 3 Cluster Compactness

Objective 1 : Cluster Compactness	
Input Dataset	*F-Measure Value*
Cancer	0.969±0.008
Iris	0.886±0.004
Liver disorder	0.928±0.002
Lungcancer	0.887±0.003
Newthyroid	0.944±0.001
Wine	0.849±0.006

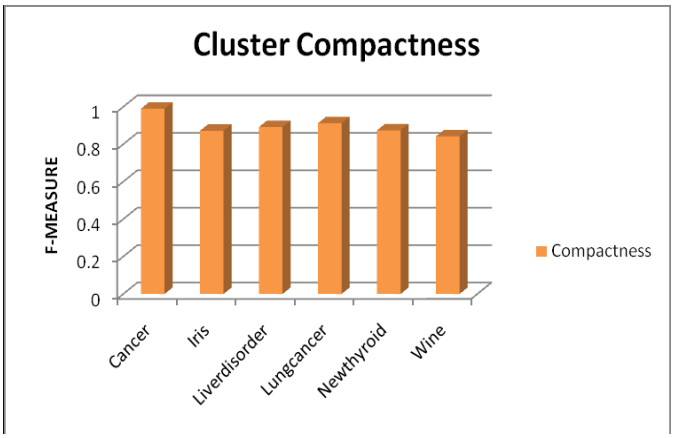

Fig.2. Graph for cluster compactness

Table 4 summarizes the F-Measure Obtained for connectedness of cluster from the seven Multiobjective clustering algorithms for the six data sets and the Fig. 3. Shows graph for cluster connectedness.
Objectve function 2 : Connectedness
Performance Evaluation Method : F-Measure

Table 4 Cluster connectedness

Objective 2 : Cluster Connectedness	
Input Dataset	*F-Measure Value*
Cancer	0.99±0.002
Iris	0.872±0.004
Liver disorder	0.892±0.006
Lungcancer	0.912±0.002
Newthyroid	0.874±0.004
Wine	0.842±0.006

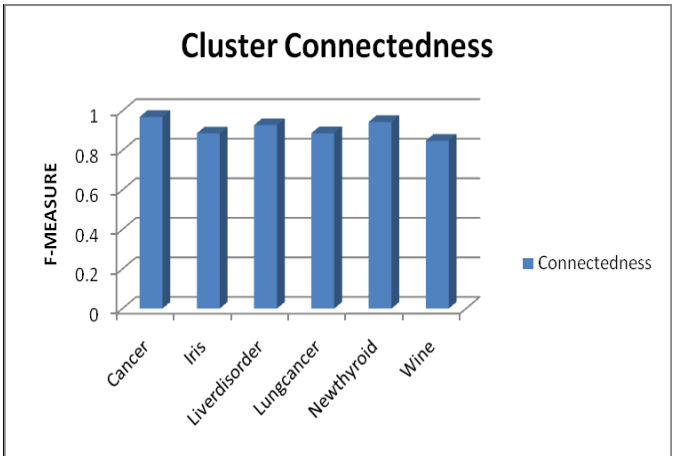

Fig.3. Cluster Connectedness

Table 5 summarizes the F-Measure Obtained for symmetry of cluster from the seven Multiobjective clustering algorithms for the six data sets and the Fig. 4. Shows graph for cluster symmetry.

Objective function 3: Symmetry
Performance Evaluation Method: F-Measure

Table 5. Cluster Symmetry

Objective 3 : Cluster Symmetry	
Input Dataset	*F-Measure Value*
Cancer	0.969±0.004
Iris	0.873±0.002
Liver disorder	0.85±0.006
Lungcancer	0.996±0.004
Newthyroid	0.883±0.008
Wine	0.862±0.002

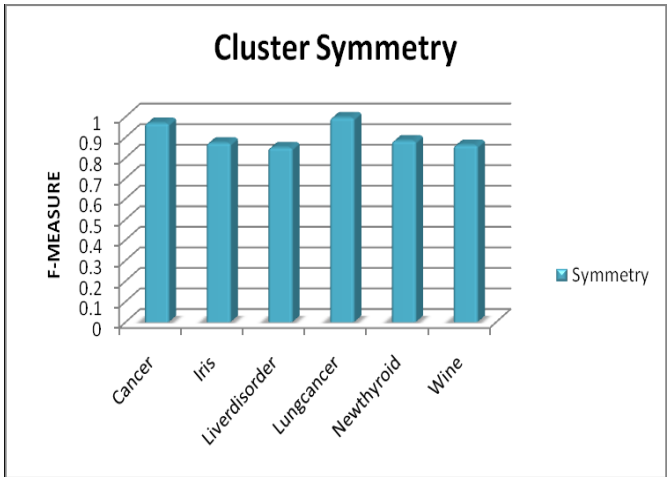

Fig. 4. Cluster Symmetry

Objective function : Multiple Objectives (Compactness, Connectedness, Symmetry)
Performance Evaluation Method : F-Measure
Table 6 summarizes the F-Measure Obtained for Compactness of cluster from the seven Multiobjective clustering algorithms for the six data sets and the Fig. 5. Shows graph for cluster compactness.

Table 6. Multiple objective values

Cluster Multiple Objectives	
Input Dataset	*F-Measure Value*
Cancer	0.990±0.008
Iris	1.0±0.002
Liver disorder	1.0± 0.002
Lungcancer	0.994±0.003
Newthyroid	0.998±0.008
Wine	0.882±0.004

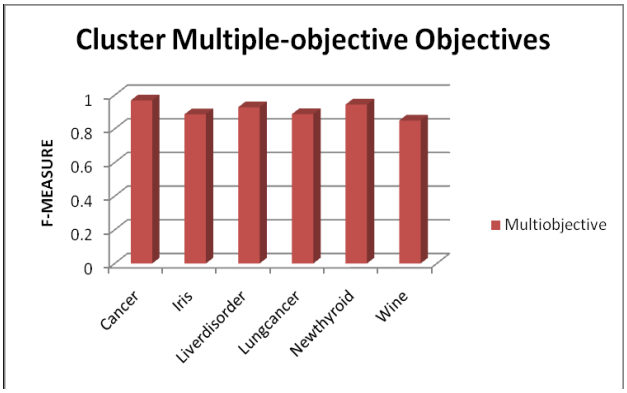

Fig.5. F-Measure for Multiple Objective

Objective function: Multiple Objectives (Compactness, Connectedness, Symmetry)
Performance Evaluation Method : F-Measure
Table 7 summarizes the F-Measure Obtained for Multiple objective functions for the six data sets and the Fig. 6. Shows comparative analysis of single objective and multiple objective functions.

Table 7. Multiple objective function

Multiple-Objective Functions (F-Measure value)				
Input Dataset	*Compactness*	*Connectedness*	*Symmetry*	*Multiple-Objectives*
Cancer	0.969±0.008	0.99±0.002	0.969±0.004	**0.990±0.008**
Iris	0.886±0.004	0.872±0.004	0.873±0.002	**1.0±0.002**
Liver disorder	0.928±0.002	0.892±0.006	0.85±0.006	**1.0± 0.002**
Lungcancer	0.887±0.003	0.912±0.002	0.996±0.004	**0.994±0.003**
Newthyroid	0.944±0.001	0.874±0.004	0.883±0.008	**0.998±0.008**
Wine	0.849±0.006	0.842±0.006	0.862±0.002	**0.882±0.004**

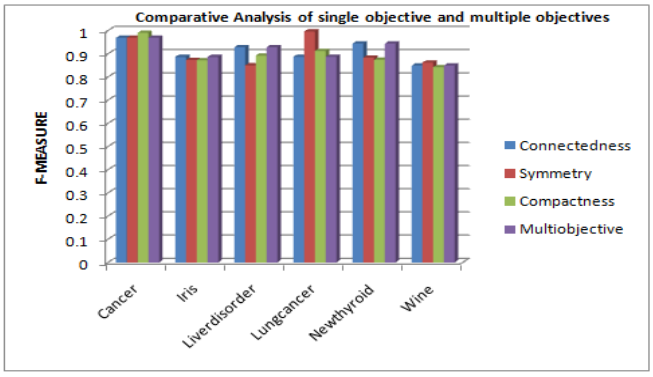

Fig. 6. Comparative analysis of single objective and multiple objectives

CONCLUSION AND FUTURE SCOPE

In this paper we proposed a data clustering technique with multiple objective functions. The k-Means algorithm will be used for initial partitioning. NSGA-II, a multi-objective algorithm will be used to optimize three objective functions. In order to achieve global optimization, the two-stage GA parameters will be applied. The proposed hybrid clustering algorithm produced quality clusters. A significant improvement is seen in terms of F-Measure over the existing hybrid algorithms. All these techniques will be used to achieve global optimization and accurate clustering. Much further work is needed to generate utility of having different and many more objectives. In Multiobjective clustering main problem is selecting the best solution from the pareto optimal front. Some new method to choose the best solution from the pareto optimal front have to be developed.

REFERENCES

[1] Sriparna Sahaa, Sanghamitra Bandyopadhyayb, "*A generalized automatic clustering algorithm in a multiobjective framework*",Department of Computer Science and Engineering, Indian Institute of Technology Patna, India, Applied Soft Computing 13 (2013) 89–108

[2] Deb, K., S. Agrawal, Amrit Pratap and T. Meyarivan (2000), "*A fast elitist non – dominated sorting genetic algorithm for multi-objective optimization: NSGA II"*. In M. S. et al. (Ed), Parallel Problem Solving From Nature – PPSN VI, Berlin, 849 –858. Springer.

[3] Hong He,Yonghong Tan," *A two-stage genetic algorithm for automatic clustering ",* International Journal of Advanced Research in Computer Science and Software Engineering, Volume 3, Issue 6, June 2013,pp-376-380

[4] S. Bandyopadhyay, S. Saha, *"A point symmetry based clustering technique for automatic evolution of clusters"*, IEEE Transactions on Knowledge and Data

[5] Xue, F.; Sanderson, A.C.; Graves, R. J. *"Pareto-based multi-objective differential evolution"*. In Proceedings of the 2003 Congress on Evolutionary Computation (CEC'2003), Canberra, Australia, 2003; Volume 2, pp. 862-869.

[6] Knowles, J. D. and D. W. Corne (1999), *"The Pareto archived evolution strategy: A new baseline algorithm for Pareto multiobjective optimization"*. In Congress on Evolutionary Computation (CEC 99), Volume 1, Piscataway , NJ,98 – 105. IEEE Press.

[7] H.C. Chou, M.C. Su, E. Lai, *"A new cluster validity measure and its application to image compression, Pattern Analysis and Applications"* 7 (July) (2004) 205–220.

[8] S. Bandyopadhyay, U. Maulik, *"Nonparametric genetic clustering: comparison of validity indices"*, IEEE Transactions On Systems, Man and Cybernetics, Part C 31 (1) (2001) 120–125.

[9] U. Maulik, S. Bandyopadhyay, "*Performance evaluation of some clustering algorithms and validity indices"*, IEEE Transactions on Pattern Analysis and Machine Intelligence 24 (12) (2002) 1650–1654.

[10] Zitzler, E. and Thiele, L.(1999). *"An evolutionary algorithm for multiobjective optimization: The strength Pareto approach"*. Technical report 43, Computer engineering and Networks Laboratory (TIK), Swiss Federal Institute of Technology (ETH) Zurich.

[11] Zitzler, E., Laumanns, M. and Thiele, L. (2001). "*SPEA 2: Improving the Strength Pareto Evolutionary algorithm".* Technical report 103, Computer engineering and Networks Laboratory (TIK), Swiss Federal Institute of Technology (ETH) Zurich.

[12] H.C. Chou, M.C. Su, E. Lai, *"A new cluster validity measure and its application to image compression, Pattern Analysis and Applications"* 7 (July) (2004) 205–220.

[13] W. Wang, Y. Zhang, *"On fuzzy cluster validity indices"*, Fuzzy Sets and Systems 158 (October (19)) (2007) 2095–2117.

[14] P.B. Helena Brás Silva, J.P. da Costa,"*A partitional clustering algorithm validated by a clustering tendency index based on graph theory"*, Pattern Recognition 39 (May (5)) (2006) 776–788.

[15] M. Kim, R. Ramakrishna, *"New indices for cluster validity assessment"*, Pattern Recognition Letters 26 (November (15)) (2005) 2353–2363.

[16] G.T. Toussaint, "*The relative neighborhood graph of a finite planar set",* PatternRecognition89912(1980)261–268.

