Multi-objective Evolutionary Algorithms for Automatic Clustering: A Comparative Study

Madhuri A. More
Computer department, PCCOE
Pune, India

***Abstract*—Evolutionary Algorithms like GA are proven to be robust in solving the Multiobjective Optimization problems. In this Paper various Multiobjective Optimization Algorithms are compared for achieving automatic Clustering. This comparison is based on various parameters to achieve Optimization. By using u NSGA-II and K-Means Clustering Algorithm a Multiobjective model is proposed.**

Keywords—Evolutionary Algorithms; Multiobjective Optimization(MOO); Clustering ; Genetic Algorithm; Symmetry.

I. INTRODUCTION

A. Clustering

Clustering is a task of Partitioning a dataset into groups such a way that object in one group is more similar to those object in other group. Generating appropriate no of cluster from given dataset is important challenge in clustering. Clustering is commonly defined as the task of finding natural groups within a data set such that data items with the same group are more similar than those within different groups[13]. A cluster is a group of objects which are similar in some way to each other and are dissimilar to the objects in other clusters. A definition of clustering is the process of grouping objects into one cluster whose members are similar or dissimilar in some way. The main issue is some data clustering algorithms may give good result with one type of data set but may fail or give poor result with other types of dataset. Generating the appropriate number of clusters is an important challenge in clustering [1].Clusters include groups in which the group members are on smaller distances. Clustering can therefore be considered as a multi-objective optimization problem. The appropriate clustering algorithm and parameter depend on the individual data set and intended use of the results. Cluster analysis is an iterative process of knowledge discovery or interactive multi-objective optimization that involves trial and failure. It is not an automatic task, until the result achieves the desired properties it will often be necessary to modify data preprocessing and model parameters. The number of clusters and the partition, are selected according to the optimal clustering tendency index value[10].

B. Multi-Objective Optimization

Multi-objective optimization is related with mathematical optimization problems, is a task of multiple criteria decision making involving multiple objective functions to be optimized simultaneously. Multi-objective optimization also called as multicriteria optimization, vector optimization, multiobjective programming, multiattribute optimization or Pareto optimization. In the single objective optimization there is only one solution, but in multiobjective optimization there is a set of solutions, called the Pareto –optimal set [14]. A multi-objective optimization problem is an optimization problem in which more than one objective functions optimized simultaneously. Multi-objective optimization is applied to many fields such as engineering, economics, science, logistics where optimal decisions need to be taken in between two or more conflicting objectives. Fig.1 shows Multi-Objective Optimization on Pareto Domain set to achieve Multiple Objective Functions. In case of nontrivial multi-objective optimization problem, there does not exist a single solution that consecutively optimizes each objective. In this situation, there exist a infinite number of Pareto optimal solutions and the objective functions are said to be conflicting. All Pareto optimal solutions are considered equally good without considering additional subjective preference information.

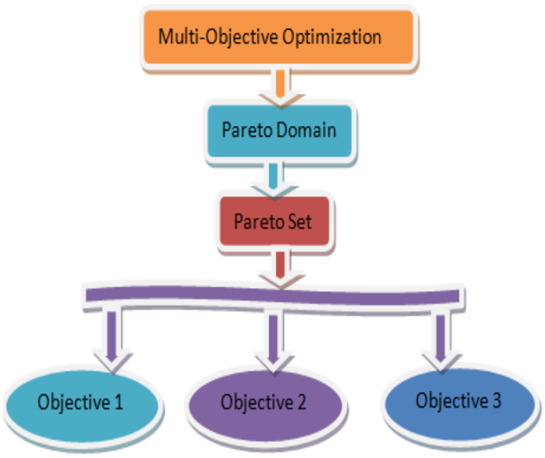

Fig.1.Multiobjective Optimization

II. MULTIOBJECTIVE EVOLUTIONARY ALGORTHMS

A. NSGAII Algorithm

K. Deb and his students [3] suggested a fast elitist Nondominated Sorting Genetic Algorithm (NSGA II)[2].In GAs, chromosomes are encoded in the form of strings [15]. In NSGA II, number of solutions that dominate solution x is calculated for each solution x . The set of solutions which are dominated by x is also calculated. The first front nondominated solutions are obtained. Let us , the set of solutions that are dominated by the solution xi is denoted by S. For each solution xi from the current front consider each solution xq from the set S. The number of solutions that dominates xq is reduced by one. After this reduction The solutions which remain nondominate, will form a separate list. Using the newly identified front as the current front this process continues. Let us consider initial population P(Q̄) is of size N. From current population P(Q̄) an offspring population Q(Q̄) of size N is created. Let us, consider the combined population [2].

$$R(\bar{Q}) = P(\bar{Q}) \cup Q(\bar{Q}).$$

Population R(Q̄) is ranked according to nondomination. By considering individuals from the fronts F1, F2, ..., New population P(Q̄+1) is formed until the population size exceeds N. According to a crowded comparison relation Solutions of the last allowed front are ranked. NSGA II uses a crowding distance parameter for density estimation for each individual. According to the crowded comparison distance Solutions of the last accepted front are ranked. Working of NSGA II is as follows. Initially Based on the nondomination a sorted random population is created[2]. Initially equal to its nondomination level each solution is assigned a fitness value. For creating an offspring population Selection, mutation and recombination are used. From the parent and offspring population a combined population is formed. According to the nondomination relation the population is sorted. For finding winner the Crowding comparison procedure is used during the population reduction phase and in the tournament selection. With a two-stage selection and mutation strategy, both selection probability and mutation probability vary with the consistence of the number of clusters in the population [5].

B. Multi-Objective Differential Evolution

MODE algorithm uses a variant of the original DE, in which to create the offspring 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.

C. Pareto Archived Evolution Strategy (PAES)

Knowles and Corne [7] have been proposed a simple evolutionary algorithm called Pareto Archived Evolution Strategy (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 [2]. The offspring is discarded if the parent dominates the offspring, and the new offspring is generated. 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 Pareto front. 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 the any member of the archive does not dominated by 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 [2].

D. Strength Pareto Evolutionary Algorithm (SPEA)

Strength Pareto Evolutionary Algorithm is described in this paper (SPEA) ([11], [9]). At every generation the algorithm maintains an external population by storing all nondominates solutions obtained so far. External population is mixed with the current population at each generation. Fitness are assigned 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. A deterministic clustering technique is used for ensuring diversity among nondominates solutions.

E. Strength Pareto Evolutionary Algorithm (SPEA 2)

Zitzler, Laumanns and Thiele [12] have proposed SPEA 2 as a variant of SPEA. SPEA2 consist of two populations. As in the Initial phase external population is empty. All nondominated solutions from current and external population are passed in the next population after the fitness evaluation [2]. Next population is fill with dominates individuals from current and external population if the number of these solutions is less than population size. Fitness assignment and a truncation operator are the main differences between SPEA and SPEA 2. From the external and current populations the fitness function is differently calculated for the solutions. In distinguish to SPEA; SPEA 2 uses a fine – grained fitness assignment strategy. This incorporates density information to differentiate between individuals having identical fitness value. As the archive size is fixed. The archive is filled up by dominates individuals whenever the number of nondominated individuals is less than the predefined archive size. The clustering technique that SPEA uses when the nondominate front exceeds the archive limit has been replaced by an alternative truncation method. The truncation method does not loose boundary points. In SPEA 2 only members of the archive participate in the mating selection process.

F. Multiobjective Genetic Algorithms

In multiobjective optimization both fitness and selection must support several objectives. Therefore, multiobjective algorithms differ from simple genetic algorithm in the way the fitness assignment and selection works. Several different variants of multiobjective algorithms have been introduced with different fitness assignment and selection strategies. Based on fitness assignment and selection strategies, multiobjective algorithms can be classified as aggregation based approach, population based approach and Pareto based approach [2]. The first multiobjective genetic algorithm, vector evaluate algorithm (VEGA) was proposed by Schaffer [8]. Afterwards, several multiobjective evolutionary algorithms were developed, such as: Multiobjective Genetic Algorithm (MOGA), Non-dominated Sorting Genetic Algorithm (NSGA), Niched Pareto Genetic Algorithm (NPGA), Weight-based Genetic Algorithm (WBGA), Strength Pareto Evolutionary Algorithm (SPEA), Random Weighted Genetic Algorithm (RWGA), Pareto-Archived Evolution Strategy (PAES), Fast Non-dominated Sorting Genetic Algorithm (NSGA-II). Generally, multiobjective genetic algorithms differ based on their fitness assignment procedure, elitism or diversification [3].

G. Adaptive Pareto Algoritm (APA)

A new algorithm for multiobjective optimization is called Adaptive Pareto Algorithm (APA). APA uses a new technique called as Adaptive Representation Evolutionary Algorithm (AREA)[4]. this technique allow each solution be encoded over a different alphabet and representation of a particular solution is not fixed. Representation is adaptive; it can be changed during the search process as effect of mutation operator. Each AREA individual is represented as a pair (x, B) where x is a string of symbols from the alphabet {0, 1, …, B-1}and B is an integer number, B ≥ 2. The standard binary encoding is generated if B = 2. The alphabet may change during the search process. APA considers a single population of individuals. Each individual is 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. The another alphabet is chosen if the parent dominates the offspring obtained in k successive mutations and the parent is represented in symbols over this alphabet. In this case the encoded solution does not change only the representation is changed. Effective and efficient diversity preserving mechanism is generated by an adaptive representation mechanism and the survival strategy.

H. Vector Evaluated Genetic Algorithm (VEGA)

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 [8]. The name is appropriate for multiobjective optimization, because the algorithm evaluates an objective vector. VEGA is a straightforward 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. In such a way each M objective functions is used to evaluate some members in the population. This algorithm emphasizes solutions which are good for individual objective functions. VEGA uses the proportional selection operator. In order to find intermediate solutions, Schaffer allowed crossover between any two solutions in the entire population. In this way, a crossover between two good solutions, each corresponding to a different objective may find offspring which are good compromised solutions between the two objectives. The mutation is applied to each individual as usual. Criticisms of VEGA [7] include the following arguments. VEGA is very simple and easy to implement, since only the selection mechanism has to be modified. One of its main advantages is that, despite its simplicity, can generate several solutions in one run of the algorithm. However, this shuffling and merging of all the subpopulations that VEGA performs corresponds to averaging the fitness components associated with each of the objectives. Since Schaffer uses proportional fitness assignment, these fitness components are in turn proportional to the objectives themselves. Therefore, the resulting expected fitness corresponds to a linear combination of the objectives where the weights depend on the distribution of the population at each generation. That is VEGA has the same problems as a previously discussed algorithm.

III. EVALUATION OF RELATED WORK

TABLE1 shows the comparison of various Multiobjective optimization algorithms. Different algorithm uses the different fitness strategy such as – bookkeeping strategy, (1+1) evolution strategy, fine grained fitness assignment strategy, elitism strategy. All the algorithms uses different clustering approach such as crowded comparison approach, aggregation based approach, truncation method, centroid based representation scheme. The algorithms are also compared on the basis of approximation, such that the NSGA-II is having good approximation as compare to the related algorithms.

TABLE1 COMPARISON OF VARIOUS MULTIOBJECTIVE ALGORITHMS

Algorithms	Fitness Strategy	Approach Used	Computational Complexity	Approximation
NSGA-II	bookkeeping strategy	crowded comparison approach	O(MN^2)	good approximation of the Pareto Optimal front
MODE	Pareto archived Evolution	centroid-based representation scheme	O(MN^3)	good approximation of the Pareto front
PAES	(1+1) evolution strategy(local search)	mutation of one offspring	O(MN^2)	Moderate approximation of the Pareto Front
SPEA	A deterministic clustering technique	clustering technique	O(MN^3)	Moderate approximation of the Pareto Front
SPEA-2	fine – grained fitness assignment strategy	truncation method	O(MN^2)	good approximation of the Pareto front
MOGA	Elitism Strategy	aggregation based approach	O(MN^2)	good approximation of the Pareto front
APA	Adaptive representation mechanism and the survival strategy	Adaptive Representation Evolutionary Algorithm (AREA) technique	O(MN)	good approximation of the Pareto front
VEGA	corresponds to a linear combination of the objectives	evaluates an objective vector	O(MN^2)	good approximation of the Pareto Optimal front

IV. PROPOSED WORK

The proposed work is for achieving data clustering using multiple objective functions. NSGA-II is popular multiobjective optimization algorithm. It will be used in proposed work to optimize the objective functions. Various objectives of clustering are given in the literature. We will focus on cluster compactness, connectedness, and symmetry. In order to achieve global optimization the evolutionary algorithms will be used. *Fig.2* shows model of proposed system.

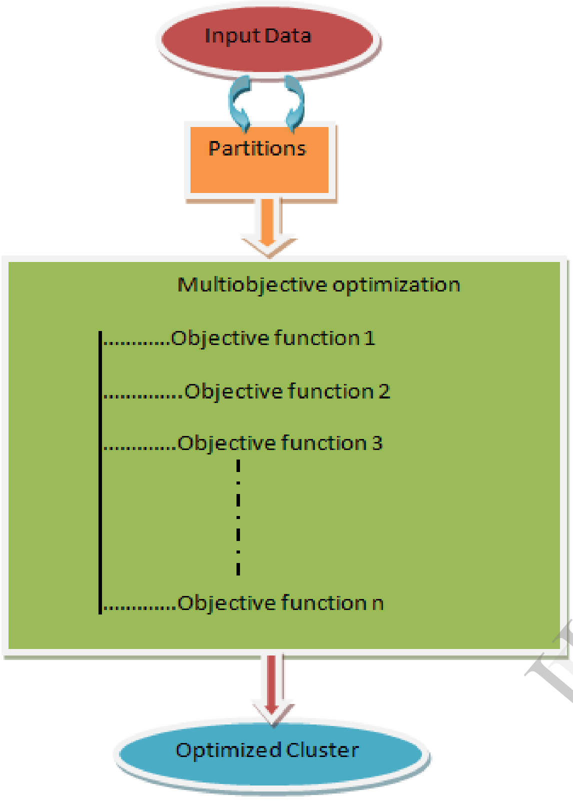
Input Data
Partitions
Multiobjective optimization
............Objective function 1
..............Objective function 2
.............Objective function 3
.............Objective function n
Optimized Cluster

Fig.2 Workflow of proposed system

V. CONCLUSION AND FUTURE WORK

A new multiobjective (MO) clustering technique will partition the data into an appropriate number of clusters. The proposed system will detect the appropriate partitioning from datasets and optimizes the multiple objective functions. Much further work is needed to generate utility of having different and many more objectives.

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]. Crina Groşan and D. Dumitrescu,"A comparison of multiobjective evolutionary algorithms" acta universitatis apulensis.

[3]. 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.

[4]. Mihaela simona Cîrciu and florin leon," comparative study of multiobjective genetic algorithms"

[5]. 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

[6]. 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.

[7]. 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.

[8]. 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.

[9]. Zitzler, E., Deb, K. and Thiele, L.(1999), Comparison of multiobjective evolutionary algorithms: empirical results. Technical report 70, Computer Engineering and Networks Laboratory (TIK), Swiss Federal Institute of Technology (ETH) Zurich.

[10]. 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.

[11]. 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.

[12]. 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.

[13]. J. Handl, J. Knowles, "Multiobjective Clustering with Automatic Determination of the Number of Clusters", 2007.

[14]. S. Bandyopadhyay, S. Saha, U. Maulik, K. Deb, "A simulated annealing based multi-objective optimization algorithm: AMOSA", IEEE Transactions on Evolutionary Computation 12 (June (3)) (2008) 269–283.

[15]. U. Maulik, S. Bandyopadhyay, "Genetic algorithm based clustering technique",Pattern Recognition 33 (2000) 1455–1465.