Journal of Applied Sciences

ISSN 1812-5654

science alert

ANSInet
an open access publisher
http://ansinet.com

Journal of Applied Sciences 14 (7): 719-722, 2014
ISSN 1812-5654 / DOI: 10.3923/jas.2014.719.722

Study on Fuzzy Clustering of Joining with Local Learning Operator in the Artificial Immune Multi-objective Optimization Algorithm

[1,2]Zhao Mengling and [1]Liu Hongwei
[1]School of Mathematics and Statistics, Xidian University, China
[2]School of Science, Xi'an University of Science and Technology, China

Abstract: As a powerful analytical tool of data mining, clustering analysis has gained considerable attention. This study first introduces the concept of artificial immune multi-objective optimization algorithm and the relevant theories of multi-objective clustering algorithm. Secondly, because the traditional clonal selection algorithm is applied to the clustering analysis, too many parameters exist, proposes a new algorithm of applying the multi-objective optimization algorithm joined with local learning operator to the multi-objective fuzzy clustering. Finally, the proposed fuzzy clustering method is applied to an artificial data set, the simulation results show that the algorithm has high clustering accuracy.

Key words: Multi-objective optimization, local learning operator, fuzzy clustering

INTRODUCTION

Data Mining (DM) is a core of knowledge discovery process in database (Bao *et al*., 2012). Cluster analysis,an important branch of data mining and an important tool of simplifying data, provides an integral technical support for DM to achieve automated information.

Optimization problem is one of the main problems in engineering practice and scientific research. The optimization problems of containing only one objective function are known as a single objective optimization problem and the optimization problems of containing two or more objective functions are known as multi-objective optimization problems. Recently, artificial immune system for solving multi-objective optimization problems has gained the interest of many researchers. The non-dominated neighbor immune algorithm (Gong *et al*., 2008) proposed by Jiao and Gong is one of the representative algorithms.

In this study, based on the multi-objective optimization algorithm of artificial immune system, we join the local learning operator for multi-object clustering.

MULTI-OBJECTIVE OPTIMIZATION

In the multi-objective optimization problems restrictions are placed on the decision variables. The optimization on one of the objects must be at the cost of the other objects and unit of each object is often not the same. So it is difficult to objectively evaluate the advantage and disadvantage of multi-objective optimization problems. In multi-objective optimization problems, one solution is maybe the best regarding one object but is not favorable of the other objects. Thus there exists a compromised solution called pareto-optimal set or non-dominated set.

Mathematical formulation of the multi-objective optimization problems (Meng *et al*., 2008; Gong *et al*., 2010): A multi-objective optimization problem with n decision variables and m object variables can be expressed as follows:

$$\begin{cases} \min \quad y = F(x) = (f_1(x), f_2(x), \cdots, f_m(x))^T \\ \text{s.t.} \quad g_i(x) \le 0, \quad i = 1,2,\cdots,q \\ \qquad h_j(x) = 0, \quad j = 1,2,\cdots,p \end{cases} \tag{1}$$

where, $x = (x_1,..., x_n) \in X \subset R^n$ is an n-dimensional decision vector from the n-dimensional decision space x, $y = (y_1,.., y_m) \in Y \subset R^n$ is an m-dimensional objective vector from the m-dimensional objective space Y. The objective function F(x) defines m mapping functions from the decision space to the objective space. $g_i(x) \le 0$ (i = 1, 2,.., q) defines q inequality constraints. $h_j(x) = 0$ (j = 1, 2,..., p), defines p inequality constraints. In this framework, several important definitions are given as follow:

- **Feasible solution:** for one $x \in X$, it is called feasible solution if it satisfies the constraint conditions (i = 1, 2,..., q) and $h_j(x) = 0$ (j = 1, 2,..., p) given above

Corresponding Author: Zhao Mengling, School of Mathematics and Statistics, Xidian University, China

- **Feasible solution set:** The set composed of all feasible solutions is called feasible solution set and is recorded as X_f, $X_f \subseteq X$
- **Pareto predominant:** Assume x_A and x_B are two feasible solutions of above multi-objective optimization problems, x_B is called pareto superior than x_A if and only if:

$$\forall_i = 1, 2,..., m, \quad f_i(x_A) \leq f_i(x_B) \tag{2}$$

$$\exists_J = 1, 2,..., m, \quad f_j(x_A) < f_j(x_B) \tag{3}$$

It is notated as $x_A > x_B$ and called x_A control x_B.

- **Pareto optimal solution:** Solution $x^* \in X_f$ is called pareto optimal solution (non dominated solution) if and only if the following condition is satisfied:

$$\neg \exists x \in X_f : x > x^* \tag{4}$$

- **Pareto optimal solution set:** Pareto optimal solution set is the set of all Pareto optimal solution:

$$P^* \triangleq \{x^* | \neg \exists x \in X_f : x > x^*\} \tag{5}$$

- **Pareto frontier:** The surface composed of all objective vectors in the Pareto optimal set P* in association with pareto optimal solutions is called Pareto frontier PF:

$$PF^* \triangleq \{F(x^*) = (f_1(x^*),..., f_m(x^*))^T | x^* \in P^*\} \tag{6}$$

Evolutionary multi-objective optimization algorithm (Bin *et al.*, 2012; Tao *et al.*, 2013): As a heuristic search algorithm, evolutionary algorithms have been successfully applied to the multi-objective optimization field. It has been developed into a relatively hot research direction Evolutionary Multi-objective Optimization (EMO).

Rosenberg (1967) recommended dealing with multi-objective optimization problems based on the evolutionary searching but the idea was not implemented. Holland (1975) proposed the genetic algorithm. Ten years later, Schaffer proposed the vector evaluated genetic algorithm, for the first time he initiated the combination of genetic algorithm and multi-objective optimization problems. Goldberg (1989) proposed a new idea for solving multi-objective optimization problems by combining the pareto theory in economics with evolutionary algorithm in his book "Genetic Algorithms in Search, Optimization and Machine Learning", which provided a significant contribution to guide the study of subsequent evolutionary multi-objective optimization algorithm.

Evaluation of the multi-objective optimization algorithm solution: To evaluate the convergence of solution and its distribution uniformity, the convergence metric proposed by Deb and Spacing metric proposed by Schott (1995) are commonly used. They are defined as below.

Convergence metric: Let $P^* = (p_1, p_2,..., p_{|P^*|})$ be the pareto optimal solution set of uniform distribution on ideal pareto frontier, $A = (a_1, a_2,..., a_{|A|})$ be the approximate pareto optimal solution set obtained by using the EMO algorithm. We can get the minimum normalized Euclidean distance d_i of the solution of distance P* for each solution a_i in set A:

$$d_i = \min_{j=1}^{|P^*|} \sqrt{\sum_{m=1}^{k} \left(\frac{f_m(a_i) - f_m(p_j)}{f_m^{max} - f_m^{min}} \right)^2} \tag{7}$$

where, f_m^{max} and f_m^{min} is the minimum and maximum of the m-th objective function in set P*. The convergence metric is defined as the average of normalized distance of all the points in set A:

$$C(A) \triangleq \frac{\sum_{i=1}^{|A|} d_i}{|A|} \tag{8}$$

The convergence metric represents the distance of algorithm between the approximate pareto optimal solution set and ideal pareto frontier. Therefore the lowerthe metric, the better the convergence of the solution and the closer it is to the ideal Pareto frontier.

Spacing metric: If the set A is the approximate pareto optimal solution set achieved by the algorithm, the spacing metric S can be defined as:

$$S \triangleq \sqrt{\frac{1}{|A|-1} \sum_{i=1}^{|A|} (\bar{d} - d_i)^2} \tag{9}$$

Where:

$$d_i = \min_j \left\{ \sum_{m=1}^{k} |f_m(a_i) - f_m(a_j)| \right\}, \quad a_i, a_j \in A, \quad i, j = 1, 2, \cdots, |A| \tag{10}$$

$$\bar{d} = \frac{1}{|A|} \sum_{i=1}^{|A|} d_i \tag{11}$$

k is the number of the objective functions. When the S value is equal to 0, it shows that the non-dominated solution is equally spaced in the objective space.

Multi-objective optimization algorithm based on artificial immune system (Wang and Li, 2007; Shang *et al*., 2012): Non-dominated Neighbor Immune Algorithm (NNIA) has been proposed by Jiao and Gong. They simulated based on the diversity of antibody symbiosis and a few antibodies activation in immune responses, selected only a few relatively isolated non-dominant individuals as active antibodies through an individual selection method based on the Non-dominant neighbor, carried out proportional cloning based on the congestion of active antibodies and adopted restructuring operation and mutation operation different from the GA for the clonal antibody group to strengthen the searching for the Pareto frontier in sparser region.

Algorithm related technical

Objective function (Siang and Khor, 2012): Antibodies adopt the real encoding based on clustering center. The algorithm optimizes both objective functions:

$$f_1 = J_m(U,V) = \sum_{i=1}^{c}\sum_{j=1}^{n}(u_{ij})^m \left\|x_j - v_i\right\|^2 \quad (12)$$

$$f_2 = S(U,V) = \frac{\sum_{i=1}^{k}\sum_{j=1}^{n}(u_{ij})^m \left\|x_j - v_i\right\|^2}{n\min_{i,j}\left\|x_j - v_i\right\|^2} \quad (13)$$

Calculation of fitness function: After algorithm iteration carried out certain steps, select the solution which maximizes the PBMF index from the approximate pareto solution set as the optimal solution:

$$PBMF = \frac{1}{c} \times \frac{E_1 \times \max_{i,j}\left\|v_i - v_j\right\|}{\sum_{i=1}^{c}\sum_{j=1}^{n}(u_{ij})^m \left\|x_j - v_i\right\|} \quad (14)$$

Joined local learning operator multi-object clustering (Gong *et al*., 2010; Liu *et al*., 2012)

Algorithm process:

- **Initialization:** Set up the algorithm parameters, set the termination conditions, generate the initial antibody population B_0, set the population size as n_D, set t = 0
- **Update the superior antibody group:** Find out the superior antibody from antibody group B_t and form a temporary superior antibody group DT_t. If the antibody group size DT_t is not greater than the superior antibody group size's upper limit n_D, let $D_t = DT_t$, otherwise, calculate the crowding distance in the antibody group DT_t, and choose n_D antibodies with bigger crowding distance to form superior antibody group D_t
- **Termination evaluation:** If $t \geq G_{max}$, the antibody group D_t can be considered as the approximate Pareto solution set. Stop the algorithm. Otherwise, let t = t+1
- **Non-dominant neighborhood selection:** If the size of antibody group D_t is not greater than the size of active population size upper limit n_D, let active population $A_t = D_t$, otherwise, calculate the crowding distance in antibodies group D_t and choose n_D antibodies of the bigger crowding distance constituting active antibody group A_t
- **Proportional clone operation:** Carry out proportional clonal operation for antibody group A_t to get antibody group C_t after cloning
- **Local learning operation:** Antibody group C_t carry out local learning operation according to certain probability (Local learning probability p_l), to get antibody group C'_t
- **Restructuring and super-mutation operation:** Carry out restructuring and super- mutation operation for antibody group C'_t to get antibody group C''_t
- **FCM iterative operation:** Carry out a step FCM iterative operation for antibody group C''_t to get antibody group C'''_t
- Merge antibody group C'''_t and D_t to get the combined antibody group B_t; go to step 2

RESULTS AND DISCUSSION

Data sets: This experiment adopted the UCI data sets and multidisciplinary synthetic data set. Where UCI data sets contains iris, wdbc, wine, glass, breastcancer. Synthetic data set contains ASD_4_2, ASD_5_2, ASD_10_2, ASD_11_2, ASD_12_2, ASD_14_2, twenty, sizes5, AD_15_2, AD_20_2.

Algorithm parameter setting and experimental results: NNIA-MOC: The fuzzy exponential index m is 2.0, the maximum number of iterations G_{max} is 100, dominant population size n_D is 100, active population size n_A is 20, clonal population size n_C is 100, mutation probability p_m is 1/l (l is the length of the antibody, algorithm adopts real number encoding based on clustering center), local learning probability p_l is 0.8, local learning intensity s is 0.3. Algorithm was ran independently 20 times on each data set. The clustering partition average accuracy and its mean square error on each data set were calculated. At the same time, the average accuracy, Adjusted Rand Index (ARI) value and Minskowski Score (MS) value of data clustering result in each data set were also given in Table 1.

Table 1: Data clustering result of average accuracy, robustness and MS value

Data set	Average accuracy	ARI	MS
ASD_4_2	0.9951(0)	0.9877(0)	0.1337(0)
ASD_5_2	0.9452(0.0047)	0.8710(0.0099)	0.4512(0.0175)
ASD_10_2	0.9949(0.0015)	0.9886(0.0033)	0.1404(0.0214)
ASD_11_2	0.9786(0.0535)	0.9896(0.0160)	0.1173(0.0664)
ASD_12_2	0.9653(0.0533)	0.9839(0.0195)	0.1449(0.0856)
ASD_14_2	0.9142(0.0836)	0.9726(0.0207)	0.2011(0.0917)
Twenty	1(0)	1(0)	0(0)
Sizes5	0.6146(0.0440)	0.4016(0.0062)	0.7233(0.0044)
AD_15_2	0.9910(0.0347)	0.9921(0.0247)	0.0556(0.1116)
AD_20_2	1(0)	1(0)	0(0)

Table 2: The active population size, cloning size and dominant population size on the performance of the algorithm

Active population size		n_C, n_D		
	size	20	50	100
n_A	5	0.9421(0.0582)	0.9503(0.0599)	0.9348(0.0651)
	10	0.9305(0.0735)	0.9621(0.0546)	0.9622(0.0553)
	20	0.9032(0.1056)	0.9215(0.0963)	0.9450(0.0647)
	50	0.9368(0.0718)	0.9241(0.0632)	0.9422(0.0697)
	100	0.8985(0.1124)	0.9071(0.0912)	0.9591(0.0763)

The experimental results in Table 1 show that NNIS-MOC algorithms can reach 100% clustering accuracy on the data set of twenty, AD_20_2. It can approximately reach 100% clustering accuracy on the data set of ASD_4_2ASD_10_2 and AD_15_2.

Algorithm parameter analysis: For data set ASD_12_2, the algorithm was ran independently under different parameter settings. The average accuracy of algorithm is given in Table 2.

Followed the settings of NNIA in conducting multi-objective function optimization, we set active population size n_A being 20, dominant population size n_D and clonal population size n_C being 100.

CONCLUSION

In this study, multi-objective optimization algorithm (NNIA) joined with the local learning operator is used for multi-objective clustering. From the simulation results one can see that the algorithm can achieve high clustering accuracy. Of course, as we already know, due to the existence of a variety of cluster validity index, selecting the appropriate objective function will play an important role on algorithm performance. A further exploration may be considered following this direction.

ACKNOWLEDGMENT

This study is supported partly by the National Natural Science Foundation of China under Gran No. 61072144 and No. 61179040.

REFERENCES

Bao, F., X. He and F. Zhao, 2012. Applying data mining to the geosciences data. Phys. Procedia, 33: 685-689.

Bin, X., Q.I. Rongbin and Q. Feng, 2012. Constrained multi-objective optimization with hybrid differential evolution and alpha constrained domination technique. Control Theory Appl., 29: 353-360.

Goldberg, D.E., 1989. Genetic Algorithms in Search Optimization and Machine Learning. Addison-Wesley, New York, USA.

Gong, M.G., L.C. Jiao, H.F. Du and L.F. Bo, 2008. Multi-Objective immune algorithm with nondominated neighbor-based selection. Evol. Comput., 16: 225-255.

Gong, M., L. Jiao and J. Yang, 2010a. Lamarckian learning in clonal selection algorithm for numerical optimization. Int. J. Artificial Intell. Tools, 19: 19-37.

Gong, M.G., L.H. Jiao and L. Zhang, 2010b. Baldwinian learning in clonal selection algorithm for optimization. Inform. Sci., 180: 1218-1236.

Holland, J.H., 1975. Adaptation in Natural and Artificial Systems. The University of Michigan Press, Ann Arbor, Michigan.

Liu, R., X. Zhang, N. Yang, Q. Lei and L. Jiao, 2012. Immunodomaince based clonal selection clustering algorithm. Applied Soft Comput., 12: 302-312.

Meng, H.Y., X.H. Zhang and S.Y. Liu, 2008. A differential evolution based on double populations for constrained multi-objective optimization problem. Chinese J. Comput., 31: 228-235.

Rosenberg, R.S., 1967. Simulation of genetic populations with biochemical properties. Ph.D. Thesis, University of Michigan, Ann Arbor, USA.

Schott, J.R., 1995. Fault tolerant design using single and multicriteria genetic algorithm optimization. M.S. Thesis, Massachusetts Institute of Technology, Cambridge, UK.

Shang, R., L. Jiao, F. Liu and W. Ma, 2012. A novel immune clonal algorithm for MO problems. IEEE Trans. Evol. Comput., 16: 35-50.

Siang, K.L.Y. and S.W. Khor, 2012. Path clustering using a mixture of dynamic time warping technique and cluster searching index method. Int. J. Intell. Inform. Process., 3: 1-10.

Tao, X.M., P. Xu, F. Liu and D.G. Zhang, 2013. Multi-objective optimization algorithm composed of particle swarm optimization and differential evolution. Comput. Simul., 30: 313-316.

Wang, J. and F. Li, 2007. A quantum-inspired immune clonal algorithm based on real-encoding. Comput. Eng., 38: 133-136.