Medical Data Clustering Based on Multi-objective Clustering Algorithm

Shilian Chen, Yingsi Tan, Junkai Guo, Yuqin He(✉), and Shuang Geng

College of Management, Shenzhen University, Shenzhen 518060, China
heyuqin2021@email.szu.edu.cn, gs@szu.edu.cn

Abstract. With the development of massive medical data, clustering algorithm becomes an effective way for medical data processing and data mining. On the one hand, it helps medical learners find effective information patterns from massive data; on the other hand, it promotes the development of medical technology and increase productivity. For traditional clustering algorithm, a single clustering index is difficult to meet people's needs of diversity and comprehensiveness. In contrast, multi-objective clustering (MOC) considers multiple objectives at the same time, and comprehensively deals with various clustering problems and standards, such as compactness, diversity of feature selection and high data dimension. Artificial bee colony algorithm (ABC) has a faster speed and embodies the idea of swarm intelligence. It imitates the optimization process of bees, and finally obtains the global optimal value. On this basis, this paper proposed a multi-objective artificial bee colony clustering algorithm (MOC-NABC) that is combined with current better-performed clustering algorithm. It takes normalized mutual information (NMI), Calinski-Harabasz (CH), Fowlkes-Mallows index (FMI) and silhouette coefficient (SC) of clustering as the final evaluation indexes. The experiment on UCI mouse protein gene dataset shows that the overall performance effect is greatly improved, e.g. compact clustering and the effective utilization of data features.

Keywords: Multi-objective optimization problem · Multi-objective artificial bee colony algorithm · Clustering algorithm · Medical data

1 Introduction

The large scale of medical informatization and the development of artificial technology enhance medical diagnosis technology. Artificial intelligence medical has a wide range of applications in medical imaging, auxiliary diagnosis, drug research and other contexts. Artificial intelligence technology can quickly find the internal relationship diagnosis method in hundreds of data set, bear large workload and work non-stop operation, which better liberates the productivity of doctors. At the same time, when the stock of data is large enough, the efficiency and practicability of artificial intelligence technology will be more extensive. Therefore, it has important significance and application prospects for the overcoming the diseases, as well as for diagnosing various physical problems of patients. In 2019, the State Council issued instructions on the use of medical data

Y. Xu et al. (Eds.): ML4CS 2022, LNCS 13657, pp. 385–399, 2023.
https://doi.org/10.1007/978-3-031-20102-8_30

[1]. Then, Ji Ping and others scholars [2] combined the guidance of State Council and existing medical health data, suggested that we need to pay more attention to artificial intelligence and improve the generalization ability for various practical needs.

Clustering algorithm is an important method in data mining, and medical data clustering is widely used in medical field. It is employed to classify patients or other medical subjects into groups by leveraging their physical data, assisting doctor to make the initial judgment. For medical data analysis, clustering algorithm not only improves the processing speed and effective use of massive data, but also contributes to discover the hidden patterns and information in the data. Therefore, it further benefits the diagnosis of biomedical phenomena. In recent years, based on the characteristics of medical data, a variety of clustering methods have been proposed. Typical examples include gene expression data analysis, genome sequence analysis, biomedical document mining and nuclear magnetic resonance image analysis.

Traditional clustering algorithms often optimize only one clustering criterion, this may not find all the clusters with different data structures [3], or the shape hidden in the subspace of the original feature space. At the same time, because of the disastrous dimensionality and poor interpretability of data, it has been a long-standing challenge to construct grouping models with high diagnostic capability and good generalization ability. Multi-objective clustering (MOC) is a type of clustering algorithm that considers simultaneous optimization of multiple objectives to integrate various clustering problems, such as closeness, diversity of solutions and other characteristics. Artificial bee colony algorithm (ABC) is proposed by Karaboga [4]. This algorithm has the characteristics of strong comprehensive ability, and maintains excellent results in jumping off local optimization, as well as showing convergence and diversity.

Many scholars have researched the field of multi-objective clustering and gained certain achievement in recent years. Hancer [5] proposed a variable-string length based multi-objective differential evolution approach for simultaneous clustering and feature selection. In the meantime, Kuo [6] suggested analyze cluster by vector updating and jumping which involves Pareto rank assignment. Dutta [7] designed context-sensitive and cluster-orient genetic operators for an unknown number of clusters to deal with continuous and categorical data.

In order to solve the problem of data dimension disaster and poor interpretability, and get rid of the traditional single clustering index, this paper proposed a novel multi-objective clustering algorithm that is combined improved bee colony algorithm with excellent clustering algorithm This algorithm takes the standard information difference of clustering, Calinski-Harabasz index, Fowlkes-Mallows index and Silhouette Coefficient as the final evaluation indexes of multi-objective algorithm, and combines four effective clustering methods: Clustering Hierarchy (AHC) [8], K-means [9], Birch [10] and Gaussian Mixed Model (GMM) [11]. Experiments on UCI mouse protein gene dataset show that the overall performance of clustering has been improved, which makes contributions to the selection of high-dimensional medical data and the clustering pattern recognition of medical data. This paper has several achievements. Firstly, we improved artificial bee colony algorithm by adding a new learning method in the stage of leading bees and scout bees. Second, this paper proposed a MOC-NABC algorithm that combined with excellent clustering algorithm. Compared with common clustering methods, it enhances the clustering accuracy of medical item datasets. Third, we combined four clustering algorithms with different swarm intelligence algorithms (BFO, PSO, ABC) and evolutionary strategies (ES) to adjust the feature selection. We found that NABC performs well in SC and CH indicators among these algorithms, and it is relatively equal in NMI and FMI indicators with the strongest comprehensive ability.

2 Methodology

2.1 Clustering Algorithms

IN terms of clustering process, two general types of basic clustering are considered: hierarchical-based clustering algorithms and partition-based clustering algorithms. This paper will focus on four clustering algorithms as the basic clustering algorithms, including Agglomerative Hierarchical Clustering (AHC), K-means, Birch and Gaussian Mixture Model (GMM) [12].

Agglomerative Hierarchical Clustering (AHC): Hierarchical methods use a series of nested groupings of the data set to cluster a set of data objects, from a single cluster to a cluster containing all individuals, and vice versa. The former is called agglomerative hierarchical clustering, and the latter is called divisive hierarchical clustering.

K-means: IN the initialization process of K-means, randomization is usually used to determine the initial number and cluster centers. The next step is to select an appropriate heuristic algorithm, and use the iterative merging process to calculate the distance metric of the points to determine the cluster center of the data merging, until the cluster is sufficiently compact and the clusters are sufficiently separated.

Birch: Birch's idea is to use a clustering feature tree to store the statistical summary of the original data, which captures important clustering information of the original data. It calculates the similarity between nodes by a certain similarity measure, sorts them from high to low, and gradually reconnects each node.

Gaussian Mixture Model (GMM): The theoretical basis of the Gaussian mixture model is on the assumption that the data obeys the Gaussian distribution as a prior probability, which can be understood as the data generated by the Gaussian distribution. By continuously increasing the number of Gaussian distributions in the model, it is possible to continuously approximate the original form of the data.

2.2 Evaluation Metrics for Clustering Algorithms

Four objective clustering factors [13] including Silhouette Coefficient, calinski-harabasz Index, Fowlkes–Mallows index and Normalized Mutual Information are considered in the proposed in the multi-objective bee colony clustering algorithm.

Silhouette Coefficient: The silhouette coefficient is a measure of cluster validity, and if the actual labels of the clustering results are unknown, it must be evaluated using the model itself. Under the similarity measure, the silhouette coefficient represents the relationship between the intra-cluster distance and the inter-cluster distance, and it is suitable for situations where the actual category information is unknown.

Calinski-Harabasz Index: Calinski-Harabasz, known as the Variance Ratio Criterion, calculates the score by evaluating the between-class variance and the within-class variance, which is defined as the ratio of the mean dispersion within a cluster to the dispersion between clusters, and scores are higher when clusters are compact and well-separated.

Fowlkes–Mallows Index: Fowlkes-Mallows score (FMI) is a comparison number between two value, using the geometric mean of precision and recall rate, which are TP (label = true and predict label = true), FP (label = true and predict label = false, or predict label = false and label = true) and FN (label = false and predict label = false). FMI scores range from 0 to 1. Higher values closer to 1 indicate good similarity between the two clusters.

Normalized Mutual Information: Standardized mutual information uses entropy as the denominator to standardize and adjust the mutual information value, so that information sets of different magnitude units can be compared and returned. Standard mutual information is often used to compare the degree of similarity between two sets of information, which can objectively evaluate the accuracy of a set partition compared with a standard partition.

2.3 Artificial Bee Colony Optimization Algorithm

Artificial Bee Colony Optimization Algorithm (ABC) is provided by Karaboga. ABC model can be used to effectively deal with multi-variable and multi-function optimization problems. ABC algorithm is very similar to many other swarm optimization algorithms, and it is also an another mature, effective and excellent theoretical application of swarm intelligence. The algorithm simulates the behavior of bees, measures the current target, constantly updates itself and completes the iteration of artificial bees' self-updating.

The main components of bee colony optimization algorithm to realize heuristic thought are as follows: honey source (solution in solution space), leading bee (the subject of solution), following bee (the main way of information exchange), and scout bee (the key of jumping out of local optimum). The process of bee searching for high-quality honey source has the following three steps.

The position of the leading bee represents the search position. At the beginning, the moving mode of the leading bee is defined by the following formula:

$$V_{id} = x_{id} + \Phi\left(x_{id} - x_{jd}\right) \tag{1}$$

where Φ is a random number evenly distributed in [0,1], determining the interference degree of the following bees in the search process. When the fitness of the new honey source V_{id} is better than x_i, the fit is calculated by using the feedback of fitness function.

If the current fit ratio is better than the old honey source position, the new honey source position will be determined to replace the old solution x_i, otherwise, x_i will be kept.

When the following bee completes a round of position update, it will get a new round of honey source information. At this time, the leading bee will dance and share information with the following bee, and the following bee will make a choice by roulette. $P(i)$. . Formula (2)describes the cumulative fitness of per iteration.

$$p_i = \frac{fit_i}{\sum_{i=1}^{Np} fit_i} \tag{2}$$

After the following bee completes the search process, if the x_i has not been updated or changed under the limit cycle, the system will determine that it is trapped in the local optimum, the task of following bees is finished after x_i is abandoned, and at the same time, it will be transformed into reconnaissance bees. The scout bee generates a new honey source by the following formula to jump out of the local solution (honey source) and enters a new cycle. L_d, U_d describes the lower and upper limit of optimization. *trial* describes Current optimization times.

$$X_i^{t=1} \begin{cases} L_d + rand(U_d - L_d), & trial \geq \lim it \\ X_i^t, & trial < \lim it \end{cases} \tag{3}$$

2.4 Enhanced Multi-objective Clustering Framework with Improved ABC

In this paper, four representative algorithms, AHC, Birch, GMM and K-means, are used in the gene dataset of mouse protein expression. In order to better measure the clustering effect, we choose four clustering indexes (SC, FMI, SC and CH) in two categories (supervised and unsupervised) to measure the clustering effect. There are 72 feature dimensions in the dataset after processing.

In this paper, 8, 16, 24, 36 and 48 feature dimensions are selected for comparison, and the rule is applied to determine the number of cluster centers for K-means and other algorithms that need to be determined, and finally 36 feature dimensions and 8 cluster centers are selected. For the parameter settings of other clustering algorithms, we choose the default values. Figure 1 shows the steps of the improved multi-objective optimization clustering algorithm. Usually, in the optimization process of ABC algorithm or other optimization algorithms, we will simply and randomly learn from other global solutions or optimal solutions to improve the particle optimization ability. In NABC, we added a new learning method in the stage of leading bee and scout bee. The object of bee learning and information transmission is no longer the global one, but the honey source and bees with the top 10% fitness value and the global optimal solution.

The following equation describes the improvement. Rand(x_j) stands for the top 10% outstanding bees, Φ_1, Φ_2 represents the speed at which bees learn from the global optimal solution and the top 10% solution.

$$V_{id} = x_{id} + \Phi_1\left(x_{gbest} - x_{id}\right) + \Phi_2\left(rand\left(x_j\right) - x_{id}\right) \tag{4}$$

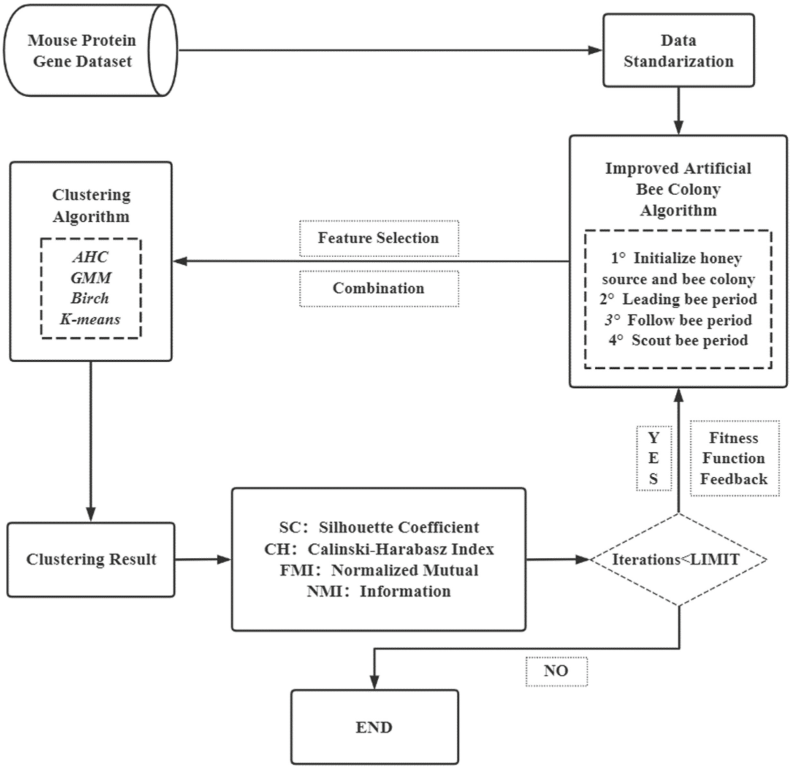

Fig. 1. The overall flow of the enhanced ABC

Under this condition, the algorithm makes use of the better solution in the whole world to make fast convergence. At the same time, it enables to promote the global search ability and avoid local optimization, therefore improve optimization efficiency.

The selected features are transferred to the clustering algorithm for model training. In the following bee stage, each following bee will use Eq. (5) to calculate its selection probability $Ps(i)$. $\sum_{i=0}^{pN} Fitness$ describes the cumulative probability of each iteration. Fitness represents the fitness value of the objective function which measured the honey source that the bee have found. Using this method to determine the object followed by the following bees ensures that the better leading bees have more probability to spread information and speed up the convergence speed. In the meantime, bees at a temporary disadvantage also have opportunities, which helps to avoid local optimal solution. Lastly, once the NABC algorithm reaches the maximum number of iterations, the method will output the best feature selection dimension and the corresponding clustering index.

$$Ps(i) = \frac{Fitness}{\sum_{i=0}^{pN} Fitness} \tag{5}$$

3 Experiments and Results

3.1 Data Set

This paper used the gene dataset of Mice Protein Expression Dataset [14] in UCI. The data compared mouse protein gene expression with the treatment effect of the Down syndrome drug “Memantine” in a biological control experiment. In today’s research on Down syndrome, the overexpression of normal chromosomal genes and chromosomal duplication have been used as a very important detection method for “Down syndrome”. Therefore, in order to detect the effect of “memantine” on the corresponding genes, the therapeutic effect of protein expression was evaluated by using mouse gene expression to evaluate drug efficacy. The dataset contains 1077 mouse samples and finally obtained 72 gene expression contents, describing eight types of mice according to characteristics such as genotype, behavior and treatment.

3.2 Experiment Settings

As shown in the below figures, the final output of the multi-objective clustering algorithm is the data feature selection and sample clustering results obtained from multiple sets of multi-index (target) training optimization. The output of each optimization algorithm is a set of Pareto solutions, in which the optimal feature combination is selected and input to the corresponding clustering algorithm to output a set of sample clustering results.

In this paper, the hardware condition is Intel (r) core (TM) i7-8550u CPU @ 1.80 ghz 2.00 ghz; The experimental environment is Python3.9. The experimental dataset is the mouse protein gene dataset from UCI. We directly use this data and extract useful information, delete a large number of missing dimensions in the data, fill in the mean of a few missing dimensions, and apply it to the experimental test of the designed multi-objective clustering algorithm.

We also selected the above multi-objective bee colony optimization algorithm and compared it with the other three common optimization algorithms, including multi-objective bacterial optimization algorithm (MOBFO) [15], multi-objective particle swarm optimization algorithm (MOPSO) [16]and multi-objective evolutionary algorithm (MOES) [17]. As the comparison algorithm in this section, the experimental parameters of the algorithm are consistent with the third section of the fourth part. The parameters were set in the clustering experiment of mouse protein gene traits (Table 1).

Table 1. Algorithm parameter setting.

Method	Value
MOC-PSO	Weight (inertia weight) = 0.8
	L1 (learning speed) = 2
	L2 (learning speed) = 2

(*continued*)

Table 1. (*continued*)

Method	Value
	Rand1 (random constant) = 0.6
	Rand2 (random constant) = 0.3
	Part_num (number of particles) = 50
	Np = feature number = 36
	Iteration = 50
MOC-NABC	Part_num (number of particles) = 50
	Fb (number of leading bee and follow bees) = 50
	Ub(upper bound of abandoning) = round(0.6*nVar*pN))
	Iteration = 50
	Np = feature number = 36
MOC-ES	N_kid (number of offspring) = 25
	pN (number of gene) = 50
	Gene_size = Feature number = 36
	Generation = 50
MOC-BFO	pN (number of bacteria) = 50
	Np = feature number = 36
	Nc (chemotaxis time) = 24
	Ns (swimming time) = 4
	Nre(reproduction time) = 2
	Ned(number of elimination) = 1
	Iteration (Nc*Nre*Ned) = 48

3.3 Experiment Results

As shown in the below figures, the final outputs of the multi-objective clustering algorithm are the data feature selection and sample clustering results obtained from multiple sets of multi-index (target) training optimization. The outputs of each optimization algorithm are a set of Pareto solutions, then select the optimal feature combination and input them into the corresponding clustering algorithm to get a set of sample clustering results.

To evaluate the final clustering performance, this paper uses 4 performance metrics for comparison. In the figures below, in order to better reflect the gradient descent clustering optimization effect, "score" is the numerical result of the index.

From the final Pareto result and the optimization curve of each index, it can be seen that NABC performs very well on SC and CH, with NABC exceeding the second BFO by 8% and NABC also performing well in CH. This shows that NABC is outstanding in the distance between cluster cohesion and cluster center in clustering effect, and in supervised learning, NABC is equal to other excellent algorithms in NMI and FMI indexes, reaching 0.5 and 0.4. At the same time, considering that NABC's optimization time and algorithm complexity are lower than BFO's, it shows greater superiority (Table 2 and Figs. 2, 3, 4 and 5).

Table 2. Experimental results of multi-objective clustering algorithm optimization.

Algorithm		SC	SC	FMI	CH
MOC_NABC	K-means	0.4136	0.5879	0.3753	**8276.9225**
	GMM	0.4861	0.5915	0.4247	5349.4911
	AHC	**0.4657**	0.5802	0.4057	7085.2524
	Birch	0.4649	**0.6370**	0.4048	**8746.6796**
MOC_ABC	K-means	0.4262	0.3671	0.3595	1499.7507
	GMM	**0.5169**	0.2072	0.4283	1401.2700
	AHC	0.4601	0.1872	0.4171	543.0246
	Birch	0.4779	0.1849	0.4142	773.2166
MOC_BFO	K-means	0.3578	0.5285	0.3578	6485.3938
	GMM	0.4629	0.5626	0.4629	6128.1491
	AHC	0.3974	0.5181	0.3974	5553.4811
	Birch	**0.4918**	0.5425	0.4106	3249.1321
MOC_PSO	K-means	0.3924	0.1763	0.3370	177.2404
	GMM	0.4586	0.1576	0.3980	158.0333
	AHC	0.4432	0.1649	0.4187	166.7836
	Birch	0.4461	0.1577	0.3923	156.7215
MOC_ES	K-means	0.3980	0.1724	0.3263	170.7654
	GMM	0.3889	0.1582	0.3727	144.5454
	AHC	0.4461	0.1561	0.3692	138.4147
	Birch	0.4355	0.1553	0.3955	138.6880

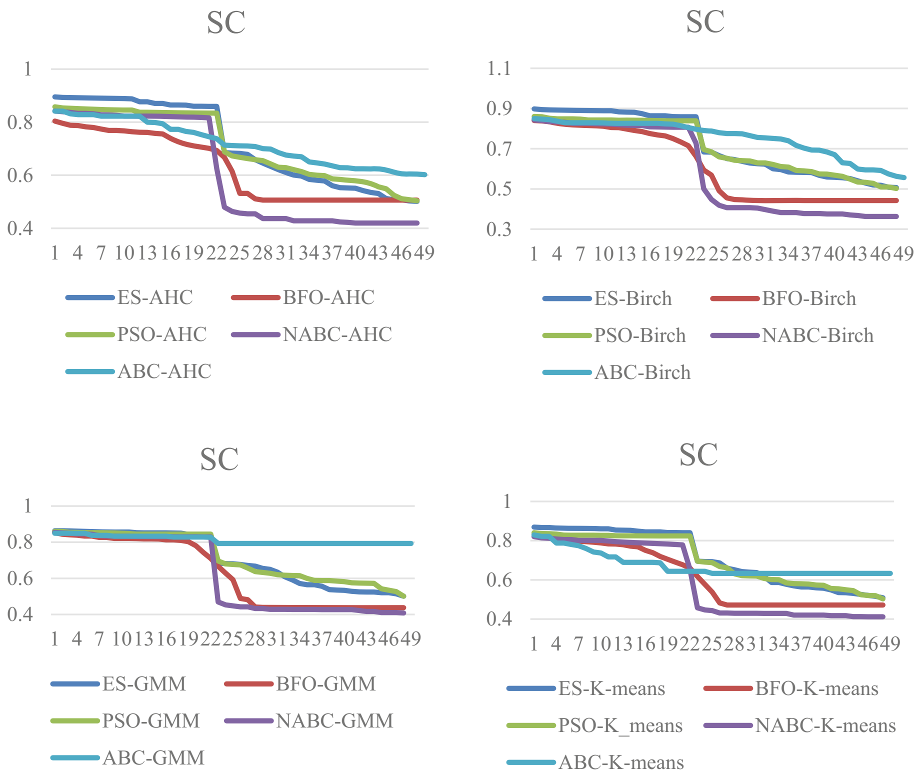

Fig. 2. The results of SC comparative experiment.

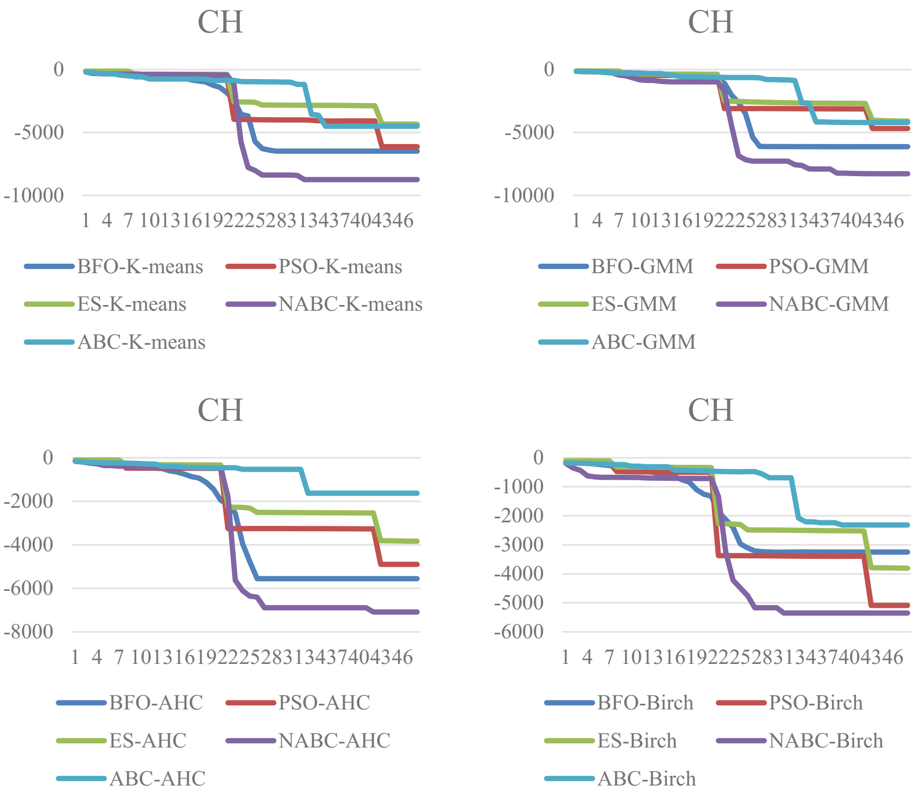

Fig. 3. CH comparative experimental results.

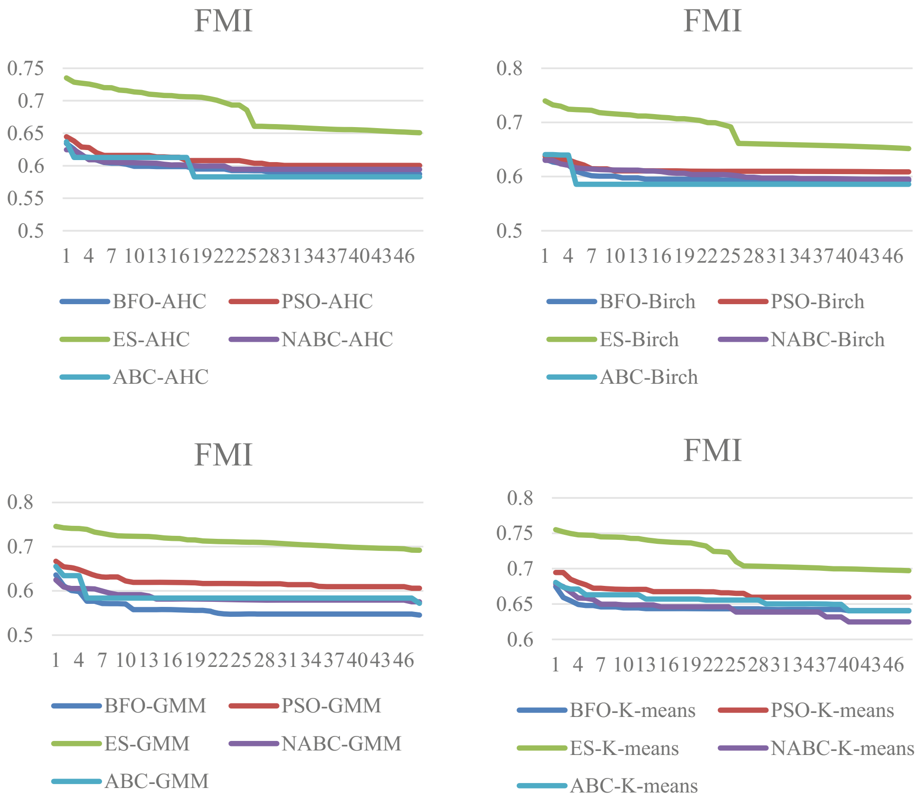

Fig. 4. FMI comparative experiment results.

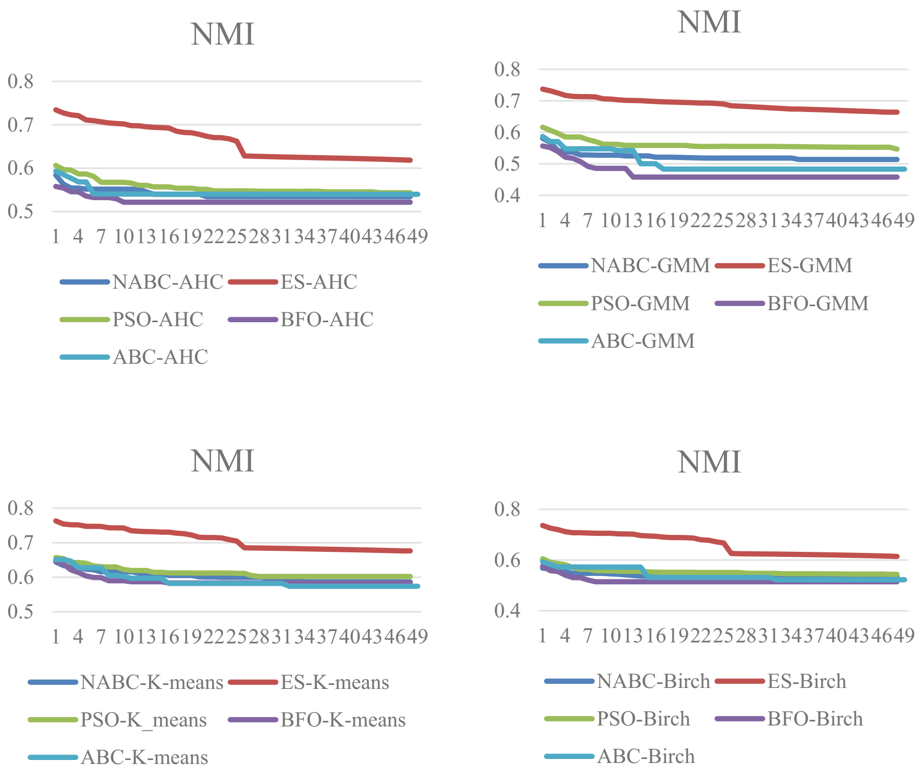

Fig. 5. NMI comparative experiment results.

4 Conclusion

In order to improve the accuracy of medical data prediction, an improved ABC (NABC) algorithm was proposed, and it is combined with four clustering algorithms and clustering indexes to form a multi-objective clustering algorithm. For ABC algorithm, the learning method of bee colony in the stage of leading bee and scout bee are improved. Besides learning the global solution, the bee colony randomly learns the top 10% solution set of the global best. In addition, in the process of following bees, selection probability is introduced to improve the performance of bee. Compared with ABC, PSO, ES, BFO and other algorithms, the performance of NABC is obviously better. This paper has several achievements. Firstly, we improved artificial bee colony algorithm by adding a new learning method in the stage of leading bees and scout bees. Second, this paper proposed a MOC-NABC algorithm that combined with excellent clustering algorithm. Compared with common clustering methods, it enhances the clustering accuracy of medical item datasets. Third, we combined four clustering algorithms with different swarm intelligence algorithms (BFO, PSO, ABC) and evolutionary strategies (ES) to adjust the feature selection. We found that NABC performs well in SC and CH indicators among these algorithms, and it is relatively equal in NMI and FMI indicators with the strongest comprehensive ability.

There are some limitations in this proposed algorithm. Firstly, after optimizing ABC's follow-up bee and scout bee stage, NABC increases additional computation, but overall, NABC's excellent effect and cost from the experiment become relatively reasonable. Secondly, the improved NABC is designed to be tested on UCI mouse protein gene dataset, and the application ability of generalization algorithm in many aspects remains to be tested. Our further research is to test and apply NABC in multiple data sets. Finally, the algorithms compared with NABC in this paper are all classic optimization algorithms. With the progress of research, there may be more excellent multi-objective clustering algorithms, and the comparison between these algorithms needs to be tried.

Acknowledgement. This study is supported Natural Science Foundation of Guangdong (2020A1515010749, 2022A1515012077), Guangdong Province Innovation Team "Intelligent Management and Interdisciplinary Innovation" (2021WCXTD002), Shenzhen Higher Education Support Plan (20200826144104001).

References

1. Office of the State Council: Guiding Opinions of the General Office of the State Council on Promoting and Regulating the Development of Health Medical Person Data Application (2021)
2. Ji, P., Zhu, D., Xie, Y.X.: Reflections on the application of scientific research sharing of health and medical data. Medicine and Philosophy **43**(1), 5–8 (2022)
3. Andreopoulos, B., An, A., Wang, X., Schroeder, M.: A roadmap of clustering algorithms: Finding a match for a biomedical application. Briefings in Bioinformatics **10**(3), 297–314 (2009)
4. Karaboga, D., Basturk, B.: Artificial bee colony (ABC) optimization algorithm for solving constrained optimization problems. Springer, International fuzzy systems association world congress (2007)
5. Hancer, E.: A new multi-objective differential evolution approach for simultaneous clustering and feature selection. Eng. Appl. Artif. Intell. **87**, 103307 (2020)
6. Kuo, R.J., Zulvia, F.E.: Multi-objective cluster analysis using a gradient evolution algorithm. Soft. Comput. **24**(15), 11545–11559 (2020)
7. Dutta, D., Sil, J., Dutta, P.: Automatic clustering by multi-objective genetic algorithm with numeric and categorical features. Expert Syst. Appl. **137**, 357–379 (2019)
8. Day, W.H.E., Edelsbrunner, H.: Efficient algorithms for agglomerative hierarchical clustering methods. J. Classif. **1**(1), 7–24 (1984)
9. MacQueen, J.: Some methods for classification and analysis of multivariate observations. In: Proceedings of the fifth Berkeley symposium on mathematical statistics and probability. Oakland, CA, USA (1967)
10. Zhang, T., Ramakrishnan, R., Livny, M.: BIRCH: an efficient data clustering method for very large databases. ACM SIGMOD Rec. **25**(2), 103–114 (1996)
11. Reynolds, D.A.: Gaussian mixture model. Encyclopedia of biometrics **41**, 659–663 (2009)
12. Jain, A.K., Dubes, R.C.: Algorithms for Clustering Data. Prentice-Hall Inc, Upper Saddle River, NJ, USA (1988)
13. Rai, P., Singh, S.: A survey of clustering techniques. International Journal of Computer Applications **7**(12), (2010)
14. Higuera, C., Gardiner, K.J., Cios, K.J.: Self-organizing feature maps identify proteins critical to learning in a mouse model of down syndrome. PLoS ONE **10**(6), e0129126 (2015)

15. Majhi, R., Panda, G., Majhi, B., Sahoo, G.: Efficient prediction of stock market indices using adaptive bacterial foraging optimization (ABFO) and BFO based techniques. Expert Syst. Appl. **36**(6), 10097–10104 (2009)
16. Zhao, L., Yang, Y.: PSO-based single multiplicative neuron model for time series prediction. Expert Syst. Appl. **36**(2), 2805–2812 (2009)
17. Kang, H.I.: A fuzzy time series prediction method using the evolutionary algorithm. In International Conference on Intelligent Computing. 530–537. Springer, Berlin, Heidelberg (2005)