Current Updates in Bioinformatics

Research Article Open Access

Microarray Gene Expression Data Clustering With Artificial Bee Colony Algorithm [Version 1, 1 Approved, 1 Approved with Reservations]

Mustafa TARIM[1] and Celal ÖZTÜRK[2]*

[1]Department of Computer Engineering, Ömer Halisdemir University, Turkey
[2]Department of Computer Engineering, Erciyes University, Turkey

***Corresponding author:** Celal ÖZTÜRK, Department of Computer Engineering, Erciyes University, Turkey, Email: celal@erciyes.edu.tr

Original Submission

Received: June 05, 2017
Accepted: June 09, 2017
Published: June 13, 2017
Last Updated: August 04, 2017

Open Peer Review Status: 1 Approved, 1 Approved with Reservations

How to cite this article: Mustafa TARIM, Celal ÖZTÜRK. Microarray Gene Expression Data Clustering With Artificial Bee Colony Algorithm [Version 1, 1 Approved, 1 Approved with Reservations], Curr Updates Bioinform (2017) 1: 2.1

Abstract

Developments in molecular biology with computer technology have created the science of bioinformatics. Many problems have been being solved in bioinformatics by computer science approaches and tools. Clustering of microarray gene expression data is one of these problems.

Artificial bee colony (ABC) algorithm, inspired by the food finding behavior of bees, is one of the most popular artificial intelligence optimization algorithms. ABC algorithm is used in the clustering problem as used in many optimization problems. In this study, we used ABC algorithm in clustering performance on microarray gene expression data.

In the study, firstly microarray gene expression data sets are selected, after that experiments are conducted. In the experiments clustering performance comparison of ABC algorithm is done with particle swarm intelligence (PSO) algorithm and some classical methods which are: K-means, partitioning around medoids (PAM) and hierarchical clustering algorithm.

Data obtained from the evaluation criteria shows that ABC algorithm is suitable for microarray gene expression clustering data and it demonstrates better performance than other algorithms in comparison.

Keywords

Clustering; Microarray Gene Expression Data; Artificial Intelligence; Artificial Bee Colony Algorithm

Introduction

With advances in molecular biology and computer science, bioinformatics has been born and has helped to explain the interrelationships of genes. Since the 1960s, many attempts have been made to study the bioinformatics field for the mystery behind the complex structures of the genes, and many problems have been identified [1]. Various solutions have been produced for the problems and these solutions have been subjected to continuous improvement. The problem of clustering of microarray gene expressions is one of them [2,3].

Artificial bee colony (ABC) algorithm is an swarm intelligence based algorithm developed by inspiration of food-seeking behavior of bees [4]. It has been used in many optimization problems and has been successful [5]. In this study, the clustering results of the gene expression data are examined by ABC algorithm.

Heuristic algorithms have been used as well as classical clustering methods for clustering of microarray gene expressions [6]. It is aimed to examine the performance of the ABC algorithm in this study. For this purpose, comparisons were made with the results obtained from various clustering methods. Among the comparison algorithms, classical methods; K-Means [7], partitioning around medoids (PAM) [8] and hierarchical clustering algorithms [9] are included. Also a heuristic method, the PSO algorithm [10] is chosen for comparison.

In the experiments, a new function for ABC and microarray gene expression clustering is coded. For comparative results, a program called CVAP [11] has been used with some improvements. The microarray data sets to be used for experiments were selected first. The selected data are known as Cho's Data [12] and Leukemia Data [13] and are frequently encountered in the literature .

Rand, FM and DBI [14-16] indices were used as evaluation criteria in clustering experiments. Also the total error rates in the cluster have been evaluated.

In the paper the results obtained from the experiments were transferred to tables and the results were interpreted in the conclusions section.

Artificial Bee Colony Algorithm

The ABC algorithm is developed by examining the food source search behavior of the bees and extracting a model from it. Despite the fact that there are a lot of bees in a bee swarm, there is no disruption in the colony or confusion in the colony due to the excellent work share between them and their ability to be self-organizing.

The ABC algorithm consists of three kinds of bees: onlooker bees, employed bees and scout bees. Every employed bee is responsible for a food source. The number of employed bees which is equal to the number of onlookers as well as equal to the number of food sources. In the event of the depletion of resources, employed bees beings become scout bees and scout bees begin to search for new resources. Thus, as a result of the depletion of resources, employed bees leave sources to provide negative feedback, while scout bees look for new sources, providing a random oscillation.

Scout bees have global research capability; employed bees and onlooker bees have regional research capability.

The ABC algorithm involves the following steps:

Step 1: Create initial food source areas

REPEAT

Step 2: Send employed bees to the food sources and calculate the amount of nectars

Step 3: Calculation of the probability values to use in the selection of the onlooker bees

Step 4: Select food sources by onlooker bees based on the calculated probability values

Step 5: Leave the abandoned food source according to the limit and produce a scout bee

UNTIL maximum cycle number is reached

One of the basic features of the ABC algorithm is that it is very flexible and simple. The algorithm simulates actual behavior of bees very closely. Developed for numerical problems but can be used in discrete problems [17].

There are various studies on data clustering with the ABC algorithm [18], but no work has yet been done for the gene clustering as far as our knowledge. The difference of the gene clustering from the standard clustering is that gene data sets has too many columns and it is difficult to generate meaningful value. For gene clustering, a new objective function must be written first. The initial set is sent according to the number of cluster centers to be created for this purpose function. The initial set consists of random solutions. According to the centers in the objective function, the distance of the gene data used is calculated. The process continues until the total distance is reduced. Clustering is completed after a certain number of iterations. The number of iterations varies according to the size and number of sets.

Microarray Technology

Microarray technology is a powerful method used to study global changes in gene expression profiles in cells and tissues. Basically, the evaluation of transcription on a genomic scale can be done in two ways. Oligonucleotide microarrays are represented by regular mosaic-shaped whole genome, obtained by photolithography method. Individual products are represented by PCR products (cDNA microarrays), obtained by spotting the DNA fragments onto the glass film by robotic spotting [19]. All of the genes of a microorganism can be micro-sized in a nail-size area, and thousands of genetic levels can be studied simul-

taneously in a single experiment [20].

Microarray technology can be used in polymorphism analysis, gene expression profiling, mutation analysis, evolutionary studies, sequence analysis, detection, development, optimization and clinical evaluation of potential therapeutic agents. Computational analysis of microarray data by looking at mRNA models also allows classification of genes as known or unknown gene classes [20].

Microarray Gene Expression Data Clustering

The process of aggregating genes with similar expression patterns in a single group is called gene expression clustering. Several clustering techniques are used in the identification of groups in the gene expression database. In the paper clustering performance comparison is done with heuristic algorithms which they are ABC algorithm, PSO algorithm and some classical methods which they are k-means, PAM and hierarchical clustering algorithm.

Hierarchical clustering algorithm is most commonly used in gene expression clustering methods. It is basically developed as a single layer artificial neural network. The nested clusters are grouped into hierarchical series to ensure that similar gene expression patterns are located close to each other. In the representation of the tree called dendogram, the relation between the genes is defined [2].

The k-means clustering algorithm is one of the most preferred algorithms in the gene clustering because of its simplicity, speed, and ease of adaptation to problems. The number of clusters (k) is given by the user. The distances between all the genes and the cluster centers are calculated. The algorithm updates the position of the centers of the clusters at each step. The final target is that the clusters to which the genes belong are close to the center point, away from the other center points [21].

The clustering studies using the PAM algorithm show very close results with the K-means algorithm, but they show more successful results than the hierarchical clustering algorithm [22].

Zhihua Du and colleagues used the PSO algorithm as a hybrid with k-means in the study they performed in gene expression clustering. While the K-means gives the algorithm a good starting point, the PSO gains speed. In this respect, an efficient hybrid algorithm has been developed [23].

Experimental Setup

In the experiments firstly Cho's Data with 386 genes and 15 time points is used. In the data set, there are 5 different clusters.

Secondly, Leukemia data set with 39 genes and 72 time points is used. This data set has 3 clusters.

Experimental work has been done on a computer with Intel Core i7 3.4 Ghz processor and 8GB RAM and the CVAP [11] was used and extensions were developed. The CVAP, developed in 2009 by Kaijun Wang et al., compares clustering methods, measures the efficiency of the algorithms works and allows users to modify their algorithms and to add data sets [11].

In the parameters of the ABC Algorithm, the colony size is set to 40, the limit 500 and the maximum number of iterations 1500. In parameters PSO algorithm, the number of particles is set to 40 and the number of iterations 1500. All of the runs are repeated 30 times for each algorithm separately and the average results of runs are given.

The method of experimental studies used in this publication is shown in Figure 1.

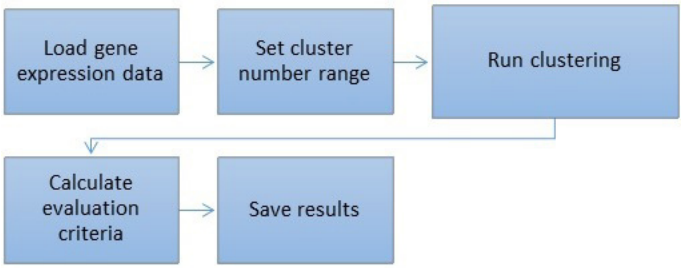

Figure 1: Method of Experimental Study.

Evaluation Criteria

There are several evaluation criteria for evaluating clustering algorithms whether they are good or bad. Some of them are used to evaluate clustering performance, and others are used to measure clustering estimation performance. Three different evaluation criteria were used in this study. These are Rand index, the FM index, and the DBI index.

The Rand index is an evaluation criterion introduced by William M. Rand in 1973. It basically compares the similarity of the values in the clusters in two separate clustering methods on an object-by-object basis. Rand index results are between 0 and 1 and if the result is 0, it means clusters do not resemble each other, and if the result is 1, it means the clusters is exactly the same [14].

FM index is developed by Fowlkes and Mallows in 1983. Although it is developed for comparison of two hierarchical clustering results, it is one of the methods frequently used in comparison of other algorithms. Despite noisy values, it makes good comparison of similarity. The clusters are examined in three parts; objects in common in clusters, objects in one cluster that are not in the other, and objects that are not in the two clusters. Bigger FM index means greater similarity [15].

DBI index is developed by Davies and Bouldin in 1979. It is used as an internal evaluation scheme of how well the simi-

larities of the quantities and content found in the clusters are. The same distance function must be used in clustering when calculating distances. The smaller index of the DBI index gives the clue that the cluster is more successful [16].

In addition to the evaluation criteria, clustering error rates are added to the comparison results.

Results

Cho's Data is known to be 5 clusters, so the performance of the algorithms in 5 clusters is important. Therefore, the performances of the algorithms in the 5 cluster are given in Table 1.

In the Table 1 the algorithms are given in rows and evaluation criteria given in columns. It shows that ABC algorithm has better result in all considering criteria. The error rate found seems a little much, but it is lower than that of other algorithms.

Table 1: Cho's Data results.

Algorithm/Index	ABC	PSO	K-Means	PAM	Hierarchical
Rand	**0,80**	0,79	0,79	0,78	0,77
FM	**0,59**	0,57	0,57	0,49	0,56
DBI	**1,27**	1,28	1,30	1,57	1,38
Error Rate(%)	**39,97**	43,52	42,05	40,05	42,22

The Figure 2 shows the result of Cho's data clustering with ABC Algorithm in Principal Component Analysis (PCA). In this PCA every color represents a cluster and some errors can seen in this figure.

Leukemia data set is known to be 3 clusters, so the performance of the algorithms in 3 clusters is important. The performances of the algorithms in the 3 cluster are given in Table 2.

In the Table 2 the algorithms are given in rows and evaluation criteria given in columns. It shows that the ABC has better results in all considering criteria. Error rates of algorithms are better than the other data set and it is very low in ABC Algorithm.

The Figure 3 shows the results of Leukemia data set clustering with ABC algorithm in PCA. Every color represents a cluster and this PCA shows that the data is almost perfect clustered.

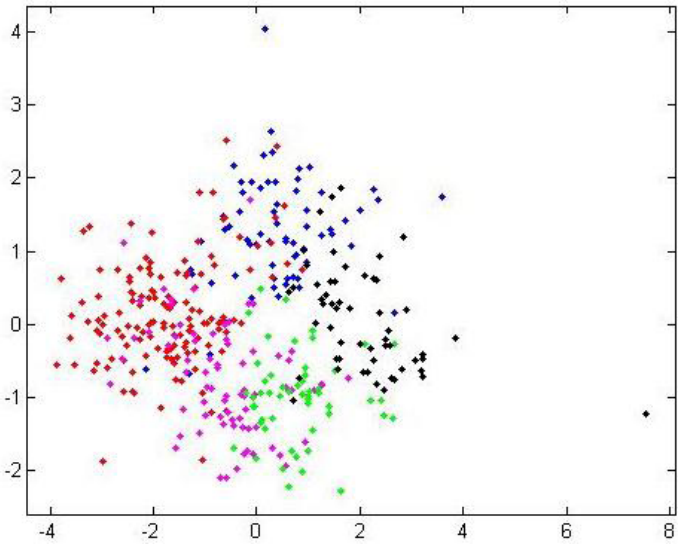

Figure 2: PCA of Cho's Data clustering with ABC Algorithm.

Table 2: Leukemia data set results.

Algorithm/Index	ABC	PSO	K-Means	PAM	Hierarchical
Rand	**0,95**	0,94	0,93	0,89	0,91
FM	**0,93**	0,92	0,91	0,84	0,86
DBI	**1,30**	1,32	1,34	1,40	1,41
Hata Oranı(%)	**3,36**	4,17	4,17	8,33	6,94

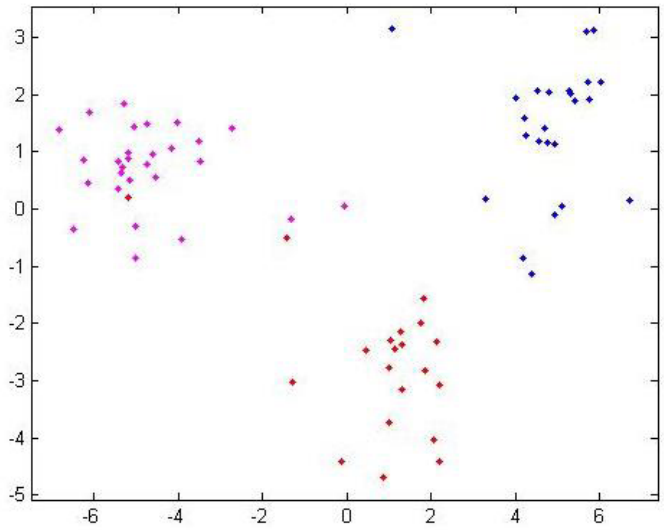

Figure 3: PCA of Leukemia data set clustering with ABC Algorithm.

Conclusions

Our experiments on two real gene expression data sets show that clustering with the ABC algorithm produces satisfactory results. It is seen that ABC algorithm is more successful than the other algorithms that are compared in this study which are PSO algorithm, K-means, PAM and hierarchical clustering algorithms. ABC algorithm can be suggested to use in clustering microarray gene expression data sets. We claim this based on 3 evaluation criteria and 4 comparison algorithms.

References

1. F Crick. Central dogma of molecular biology. Nature. 1970; 227: 561-563.
2. MB Eisen, PT Spellman, PO Brown, D Botstein. Cluster analysis and display of genome-wide expression patterns. Proc Natl Acad Sci USA. 1998; 95: 14863–14868.
3. P Tamayo, Donna Slonim, Jill Mesirov, Qing Zhu, Sutisak Kitareewan, et al. Interpreting patterns of gene expression with self-organizing maps: methods and application to hematopoietic differentiation. Proc. Natl. Acad. Sci. USA. 1999; 96: 2907-2912.
4. D Karaboga, B Basturk. A powerful and efficient algorithm for numerical function optimization: Artificial bee colony (ABC) algorithm. J. Glob. Optim. 2007; 39: 459-471.
5. D Karaboga, B Basturk. On the performance of artificial bee colony (ABC) algorithm. Appl. Soft Comput. J. 2008; 8: 687–697.
6. AK Jain, MN Murty, PJ Flynn. Data clustering: a review. ACM Comput. Surv. 1999; 31: 264–323.
7. DR Cox. Note on Grouping. J. Am. Stat. Assoc. 1957; 52: 543–547.
8. L Kaufman, PJ Rousseeuw. Clustering by means of medoids. Statistical Data Analysis Based on the L 1-Norm and Related Methods. First International Conference. 1987; 405–416.
9. P Giudici, S Figini. Applied Data Mining for Business and Industry. 2009.
10. J Kennedy, R Eberhart. Particle swarm optimization. Neural Networks, 1995. Proceedings., IEEE Int. Conf. 1995; 4: 1942–1948.
11. K Wang, B Wang, L Peng. CVAP: Validation for Cluster Analyses. Data Sci. J. 2009; 8: 88–93.
12. RJ Cho, Campbell MJ, Winzeler EA, Steinmetz L, Conway A, et al. A genome-wide transcriptional analysis of the mitotic cell cycle. Mol. Cell. 1998; 2: 65–73.
13. EP Xing, RM Karp. CLIFF: clustering of high-dimensional microarray data via iterative feature filtering using normalized cuts. Bioinformatics. 2001; 17: S306–S315.
14. WM Rand. Objective Criteria for the Evaluation of Clustering Methods. J. Am. Stat. Assoc. 1971; 66: 846–850.
15. EB Fowlkes, CL Mallows. A Method for Comparing Two Hierarchical Clusterings. J. Am. Stat. Assoc. 1983; 78: 553.
16. DL Davies, DW Bouldin. A cluster separation measure. IEEE Trans. Pattern Anal. Mach. Intell. 1979; 1: 224–227.
17. D Karaboga, B Akay. A comparative study of Artificial Bee Colony algorithm. Appl. Math. Comput. 2009; 214: 108–132.
18. D Karaboga, C Ozturk. A novel clustering approach: Artificial Bee Colony (ABC) algorithm. Appl. Soft Comput. J. 2011; 11: 652–657.
19. M Bedn. DNA microarray technology and application. Med Sci Monit. 2000; 6: 796–800.
20. MJ Heller. DNA Microarray Technology: Devices, Systems, and Applications. Annu. Rev. Biomed. Eng. 2002; 4: 129-153.
21. S Tavazoie, J D Hughes, M J Campbell, R J Cho, G M Church. Systematic determination of genetic network architecture. Nat. Genet. 1999; 22: 281-285.
22. G Chen, S Jaradat, N Banerjee, T Tanaka, M Ko, et al. Evaluation and comparison of clustering algorithms in analyzing es cell gene expression data. Stat. Sin. 2002; 12: 241-262.
23. Z Du, Y Wang, Z Ji. PK-means: A new algorithm for gene clustering. Comput. Biol. Chem. 2008; 32: 243–247.