International Journal of Scientific Engineering and Applied Science (IJSEAS) – Volume-3, Issue-5, May 2017
ISSN: 2395-3470
www.ijseas.com

Data Clustering Using Improved Genetic Algorithm

Mohammadjavad Hosseinpoor[1], Ehsan Tavakoli[2], Mahmoud Fotouh Estahbanati[3]
[1] Member of faculty in Department of Computer Engineering and Member of Young researchers and elite club, Estahban Branch UniversityEstahban, Iran
[2,3] Department of Computer Engineering and Member of Young researchers and elite club, Estahban Branch , Islamic Azad University Estahban Branch Estahban, Iran

Abstract
These Clustering is one of the most important methods in data mining field which groups data into different categories to retrieve useful information from initial data set. Clustering works with first randomly selecting the clusters' centers and then grouping the data around these centers. Innovative algorithms are heuristic algorithms used to optimize the clustering issues. To mitigate this problem, in this article, we introduce improved genetic algorithms which has been used for data collections in UCI repository. The results of the study on comparing our approach with the genetic algorithm shows effectiveness of our approach in generating quality results.
Keywords: *Data clustering, Data mining, IGA.*

1. Introduction

Clustering is an unsupervised learning method used to classify data in a variety of fields, such as data mining [1], compression [2] and pattern recognition [3]. In the first step of a typical clustering process, the K numbers of clusters' centers are selected randomly from the data set. Selecting these centers accurately and consciously can be very influential on the final results. Next, the remaining data are grouped around these centers. Genetic algorithms are kind of the computational models that are based on biological evolution [4][5]. Genetic algorithms are evolutionary algorithms and have a population of generation based on the optimization ultra-heuristic algorithms [6]. Genetic Algorithm [7] first proposed and has been developed by John Holland in 1970. In the following we discussed about the examples which are based on the genetic algorithms. Two of these examples have used the genetic algorithms in feature extraction tasks [8] and [9]. In addition, the article [10] has used genetic algorithm to discover information from databases by integrating it with neural network. On the other hand, articles [11] and [2] introduced fuzzy models based on genetic algorithms. As another example, we can point to using the genetic algorithm in developing facial recognition systems and discovering the laws from biological data [3-5]. The last but not the least is clustering in data mining [6].

Generally, genetic algorithm is an iterative process which produces new population of older population. Each chromosome thread in the population is provided in a binary form. This algorithm includes 3 genetic operations: Selection, Crossover and Mutation. These operations are applied on the initial population of chromosomes threads to produce new population. The nature of being an iterative process leads these algorithms to continually improve the quality of the results.

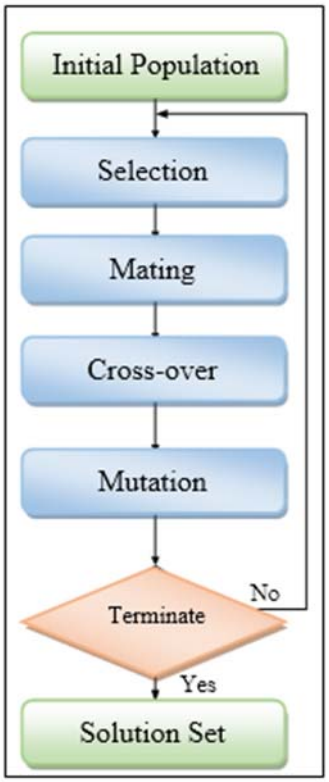

Figure 1: The genetic algorithm.

Finally, the process will end up with finding the o`ptimal solution. Figure 1 shows how the genetic

algorithm repeatedly finds the answer to a problem. This process begins with an initial population and then, as mentioned before, the three basic operations are iteratively applied until the optimal response is found.
The three basic operations are as following:

•Selection: In this step, a part of the existing population is selected to generate a new population. Selection criteria is based on the output value of the fitness function and the populations which have adequate levels of fitness are selected to be the inputs for the cross over step. Fitness function is a heuristic function depends on the problem at hand and measures and quality of the answers.

•Cross-over: Next, the new generation of parents are generated in this step. To this end, a point is randomly selected along the chromosome and chromosomal genes after this point are replaced with other chromosomal genes. There are different methods for selecting the chromosomes before the cross-over operation. Some of these examples are the Roulette Wheel, Rank, Competitive Selection and etc. Many of cross over methods are used according to the selection of this point throughout the chromosome, such as 1-point, two-point, separating, cutting and etc [7].

•Mutation: This operation provides random modes for answers and creates genetic variation between the two generations. In the mutation, an arbitrary bit (Bit) is transferred from its initial position. Examples of different types of mutations are bit string mutations, Gaussian mutations, non-uniform mutations and [8]. The mutation operation prevents the creation of local minimum by preventing the generation of similar populations of chromosomes.

2. Improved Genetic Algorithm

As discussed in previous section, genetic algorithm begins with a random population of chromosomes and then repeatedly finds the optimal solution to the problem. In the improved genetic algorithm, the first step (i.e. the random selection of initial population) is performed using three different methods: randomized, competitive and roulette wheel. Selecting of these three methods itself is a randomly process. To improve clustering results in this algorithm, the mutation is occurred on each child immediately after generating them. In the case if the value of the objective function of the mutant child was better than the worst member of the population, that member will be replaced with mutant child. This will influence on the future parent's choices and reduce the value of the objective function in comparison with basic genetic algorithms.

1. Genetic Algorithm
2. begin
3. Choose initial population
4. repeat
5. Evaluate the individual fitness of a certain proportion of the population
6. Select pairs of best-ranking individuals to reproduce with three methods (Random, Roulette well, Tournament)
7. Apply crossover operator and mutation operator on every offspring
8. Apply mutation operator on random chromosome
9. until terminating condition
10. end

Figure 2: The soducode of improve genetic algorithm

Random Selection
In this case, an element is selected randomly.
Roulette Selection
In this case, the element that has the greater value of fitness function, is selected. In fact, we assign a cumulative probability to each element according to its fitness value and identify its selection chance using this probability.
Tournament Selection
A subset of the attributes of a population are selected and its members are competing, and finally only one attribute of each category will be selected for generation.

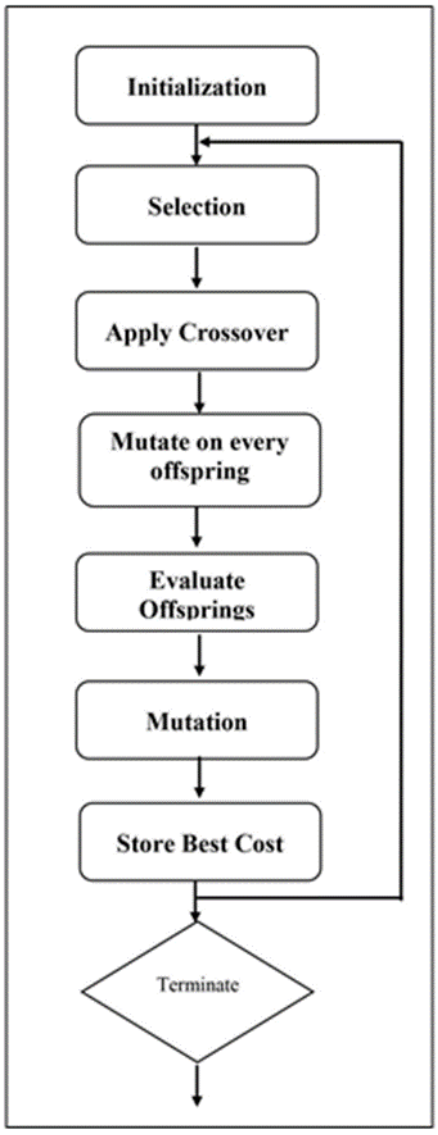

Figure 3: The improved genetic algorithm.

3. Experimental

Assessment was performed on the five available data sets from the UCI repository including Iris, Breast Cancer, SAHeart, and Galaxy [8]. The following table summarizes the characteristics of these categories. Table columns define sample size, characteristics and classes for each of the categories.
In impalement of IGA:

- Population: 100
- Repeat the inner loop: 50
- Total runs: 100

Dataset	Classes	Attributes	Instance
Iris	3	4	150
Breast Cancer	2	9	683
SAHeart	2	9	462
Galaxy	7	4	323

Table 1: The data sets.

The results of IGA are compared with genetic algorithms. This study shows that the results are competitive with the genetic algorithm. Table 2 shows the results of comparison of the total distance within the cluster.

Table 2: The comparison results of the total inner clusters distances

Algorithm / **Datasets**	**GA**	**IGA**
Iris	97.7986	**97.0546**
Breast Cancer	1193.1163	**1119.3879**
SAHeart	1230.7362	**1157.0813**
Galaxy	237.5852	**214.2524**

Figure 4 shows the comparative for IGA algorithm for the total distance within the cluster and show that this value in IGA algorithm is competitive with genetic algorithm.

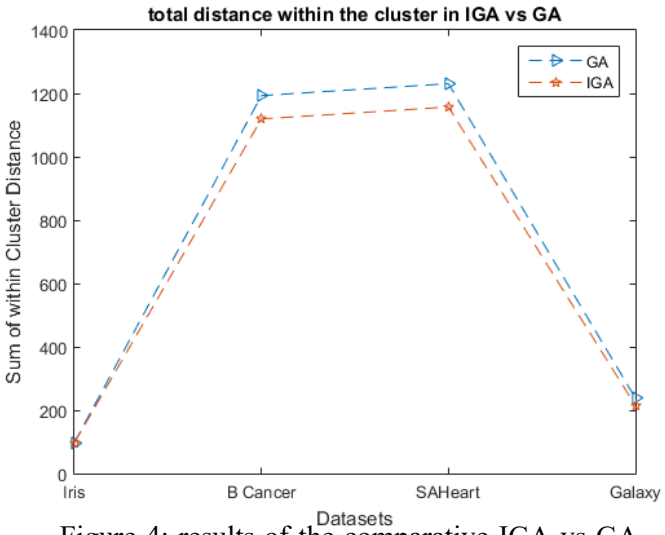

Figure 4: results of the comparative IGA vs GA in 4 datasets.

4. Conclusion

In this paper, we propose and implement IGA algorithm for clustering and we assessed that how the mutation on generated children in each generation could improve the genetic algorithm results. As the future works, first, the implementation of this algorithm on other data sets available on the UCI repository can be interesting. Secondly, the proposed method can be used to solve optimization problems and data mining in various fields such as clustering of data.

References

[1] C. Pizzuti and D. Talia, "P-autoclass: scalable parallel clustering for mining large data sets", IEEE Transactions on Knowledge and Data Engineering, vol. 15, no. 3, pp. 629-641, 2003.

[2] J. Marr, "Comparison of several clustering algorithms for data rate compression of lpc parameters", in Acoustics, Speech, and Signal Processing, IEEE International Conference on ICASSP'81., vol. 6,pp.964-966, IEEE, 1981.

[3] A. K. Wong and G. C. Li, "Simultaneous pattern and data clustering for pattern cluster analysis", IEEE Transactions on Knowledge and Data Engineering, vol. 20, no. 7, pp. 911-923, 2008.

[4] S. Russell, P. Norvig, and A. Intelligence, "A modern approach", Arti cial Intelligence. Prentice-Hall, Egnlewood Cli s, vol. 25, p. 27, 1995.

[5] D. E. Golberg, "Genetic algorithms in search, optimization, and machine learning. addion wesley",Reading, 1989.

[6] M. Pei, E. Goodman, and W. Punch, "Feature extraction using genetic algorithms", in Proceedings of the 1st International Symposium on Intelligent Data Engineering and Learning, IDEAL, vol. 98, pp. 371-384, 1998.

[7] B. G. Kermani, M. W. White, and H. T. Nagle, "Feature extraction by genetic algorithms for neural networks in breast cancer classi cation", in Engineering in Medicine and Biology Society, 1995., IEEE 17th Annual Conference, vol. 1, pp. 831-832, IEEE, 1995.

[8] A. Kannan, G. Q. Maguire, A. Sharma, and P. Schoo, "Genetic algorithm based feature selection algorithm for e ective intrusion detection in cloud networks", in Data Mining Workshops (ICDMW), 2012 IEEE 12th International Conference on, pp. 416-423, IEEE, 2012.

[9] Hosseinpoor.MJ., Kazemi.H, 2013, "Present a New Middleware to Control and Management Databases in Distributed Environment", International Journal of Emerging Technology and Advanced Engineering, ISSN 2250-2459, Volume 3, Issue 5, PP: 547-552.

[10] Z. Yuanhui, L. Yuchang, and S. Chunyi, "Combining neural network, genetic algorithm and symbolic learning approach to discover knowledge from databases", in Systems, Man, and Cybernetics, 1997. Computational Cybernetics and Simulation., 1997 IEEE International Conference on, vol. 5, pp. 4388- 4393, IEEE, 1997.

[11] P. Kr•omer, J. Platos, V. Snasel, and A. Abraham, "Fuzzy classi cation by evolutionary algorithms", in

Systems, Man, and Cybernetics (SMC), 2011 IEEE International Conference on, pp. 313-318, IEEE, 2011.

[12] A. Bozorgnia, S. H. M. Zargar, and M. H. Yaghmaee, "Fuzzy improved genetic k-means algorithm", inElectrical Engineering (ICEE), 2011 19th Iranian Conference on, pp. 1-6, IEEE, 2011.