The Journal of Mathematics and Computer Science

Available online at
http://www.TJMCS.com

The Journal of Mathematics and Computer Science Vol .2 No.2 (2011) 329-336

Representing the New Model for Improving K-Means Clustering Algorithm based on Genetic Algorithm

Rouhollah Maghsoudi[1,*], Arash Ghorbannia Delavar[2], Somayye Hoseyny[3], Rahmatollah Asgari[4], Yaghub Heidari[5]

Department of Computer, Nour Branch, Islamic Azad University, Nour, Iran
Payame Noor University, Tehran, Iran, a_ghorbannia@pnu.ac.ir
Payame Noor University of Shahrerey, sce.hoseyny@gmail.com
Islamic Azad University of Semnan, r_askaree@yahoo.com
Department of Electrical, Nour Branch, Islamic Azad University, Nour, Iran

Received: September 2010, Revised: December 2010
Online Publication: January 2011

Abstract

Data clustering into appropriate classes and categories is one of the important topic in pattern recognition. It is very good and very efficient that the number of data which misclassified is minimized or in other words data that classified in each class has been possible as much possible similarity together. In this article at the first, a fundamental method of data clustering which named **K-Means Clustering** was expressed and then with genetic algorithm , our proposal model that we named it GA-Clustering for improving K-Means method has been introduced. Finally, the said model was examined on some of the well-known data set. Results show that our method clusters data better than traditional K-Means Clustering algorithm significantly.

[1,*] Corresponding author:
E-mail address: rcemaghsoudy@yahoo.com (Phone: +98 122 625 4590)
[2] Payame Noor University, Tehran, Iran
[3] Computer Department of Payame Noor University of Shahrerey
[4] Mechanical Department of Islamic Azad University of Semnan
[5] Electrical Depertment of Islamic Azad University of Nour (Phone: +98 122 625 4590)

Keywords: Pattern Recognition, Clustering, K-Means, Genetic Algorithm.

1. Introduction

At the first it's necessary to explain about clustering. This method is an instrument for unsupervised classification of data sets. In fact, clustering is an unsupervised separating technique that get data as vectors in multi-dimensional search space and then put them in categories or clusters. With this assumption that patterns of a cluster have a similarity mean while the patterns of different clusters don't similar together. Therefore we must find a similarity measurement for some of data such that these patterns set in an identical cluster. One similarity measurement can be euqlidician distance of X and Y patterns that defined as below:

$D = || x - z ||$

If this distance becomes fewer then similarity degree will be superior.

Clustering in a N-dimensional search space is same as dividing process of N-point set into K subset or subspace (cluster or category) accordingly similarity or unsimilarity degree of data. Let's go show these N points {x1,x2,…, x*n*} become as S that it includes K clusters $C_1,C_2,\ldots,C_k$, then

$\mathbf{C_i \neq 0}$ **for i=1,…,k,**

$\mathbf{C_i \cap C_j = 0}$ **for i=1,…,k , j=1,…,k and i≠j**

And $\cup^{k}_{i=1} C_{i=s}$

In this paper K-Means clustering method will be discussed firstly.

2. K-Means Algorithm

As mentioned above, this method is one of the clustering techniques that represented by Mc. Queen in 1967. K-Means can be summarized in some steps totally:

Step 1: we select K points as cluster centers $Z_1,Z_2,Z_3,\ldots,Z_k$ among N data points randomly.

Step 2: we allocate X_i,i=1,2,…,n to hypothetical cluster C_j if and only if

$|| x_i - z_j || < || x_i - z_p ||$, P = 1,2,…,K, and j ≠ p.-

Step 3: the new centers of clusters $Z_1,Z_2,Z_3,\ldots,Z_k$ are calculated as below:

$Z_i^* = 1/\tilde{n}_i \quad \sum x_{j\,,\,i=1,2,\ldots,k}$

$x_j \in C_i$

That $\tilde{n}_i$ is the elements which belong to cluster C_j.

Step 4: if $z_i = z_i$, i = 1,2,…,K then algorithm finishes, otherwise continue from step 2.

In conditions that process doesn't terminate in step4 naturally, it's better algorithm repeat as maximum predefined iterations.

Expressed statements was summation of K-Means clustering algorithm. But next method or it's better to say our method is clustering process using genetic algorithms. Of course our technique is composition of K-Means and GA together.

3. Genetic Algorithm

Genetic algorithms are improvement techniques and some methods of accident at searching who are produced from concepts of natural selection recommendation and evolutional trends. This Darwin's evolutional recommendation is so display, in nature who can believe that has more competence for life and others will die in competition. Indeed , genetic algorithms are resolution systems base of computer and they use calculation models in some basic elements of gradual evolution in design and performance.

General element on genetic algorithms is applying an accidental searching instead of a certain and defined algorithm. Genetic algorithm is a member of more extensive family so called evolutional algorithms.

Extensive groups in evolutional algorithms are :

**Genetic algorithms*

**Evolutional programming*

**Evolutional strategies*

**Classification systems*

**Genetic programming*

All of them are identical in a basic meaning and simulation of evolution has done with selection. Mutation and reproduction trends. Genetic algorithms have structures that conclude with selection rules and other process.

Any one in valuational population receives fitness from environment. Reproduction is happen on individual who has more fitness. Mutation and crossover are practice on individuals and change them with a revelatory dependent.

General structure of a genetic algorithm is defined:

1.3. A seudo-code for Genetic Algorithm

In this part we display a general semi-code for genetic algorithms and their processes.

In this section we try to give an absolutely view of base and processes of genetic algorithms. Later we describe this process generally specific semi-code for genetic algorithms is:

1. **Representing and encoding:** Defines a genetic show from specific system.
2. **Start:** Produces an accidental population containing n chromosome.
3. **Fitness:** Assesses fitness of each chromosome in population and selects among them more value able chromosome so call particular initiative population for genetic algorithm.
4. **New population:** Apply some processes in following way to create new population:

4.1 **Selection:** Selects 2 chromosomes base of their fitness in population.

4.2 **Crossover:** Cross these 2 certain chromosomes and produces new children and adds them to the new population.

4.3 **Mutation:** With a mutation probability, it changes parent chromo some series and adds this mutant chromo some to the new population.

4.4 **Accepting:** Fits these new chromosomes and selects them if they have move competence than initiative population; if they are not deletes them from new chromosome lists.

Duty of GA is finding proper cluster centers Z_1 , Z_2 , Z_3 ,..., Z_k in such away that clustering standard of M can be minimize.

5. Replace: in this phase the new population chromosomes are replaced which chromosomes of old population that have fewer values rather than them and current population is known as primary population.
6. Test: if termination condition is satisfied then algorithm stops and selects best solution among population.
7. Loop: jump to step 4.

If we want to display a flowchart of genetic algorithm, will be as you see in below:

- Representation and encoding of problem
- Start
- Fitness
- Creation of new population
- Selection
- Crossover
- Mutation
- Accepting
- Replace
- Test
- Loop

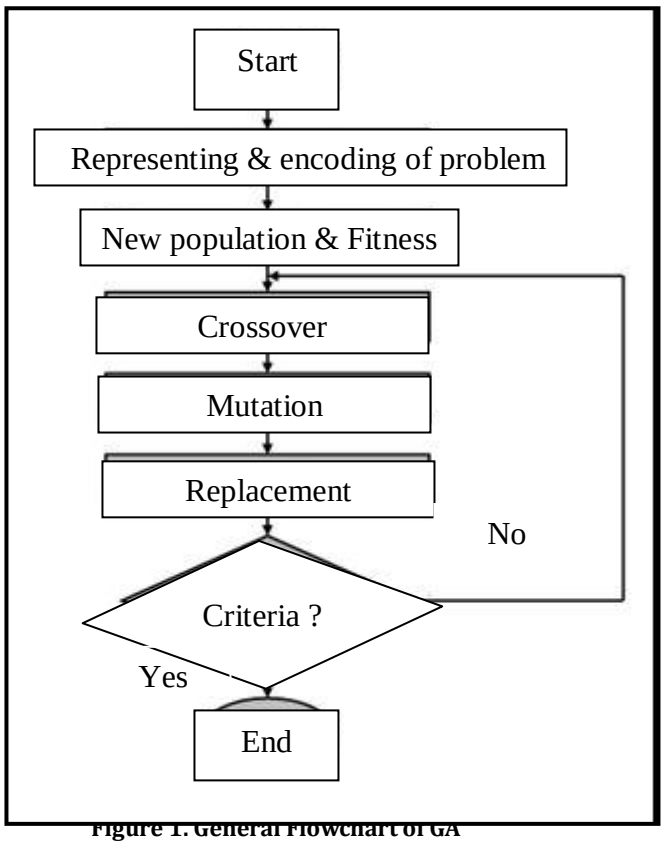

Figure 1. General Flowchart of GA
Figure 1. General Flowchart of GA

4. Clustering using GA

In clustering using GA defines a measurement naming μ as below:

$\mu(C_1,C_2,...,C_k) = \sum \sum ||X_j - Z_i||$

Duty of GA is finding proper cluster centers Z_1 , Z_2 , Z_3 ,..., Z_k in such away that clustering standard of M can be minimize.

1.4. Displaying series

Each series is a collection of real numbers that show K cluster centers.
In N-D environment, length of each chromosome is identify wit N*K.

(N_1 N_2 N_k)

Center of c_1 center of c_2 center of c_k

For example, assume that N=4 , K=3 (Iris) , this means space of question is 4D and viewable clusters are 3. So chromosomes are in this form:

1.5000 3.5000 5.2000 0.4000

1.9000 3.8000 5.1000 0.4000

1.6000 3.4000 5.0000 0,2491

That showing 3 cluster centers,

0.4000 5.1000 3.8000 1.9000

1.5000 3.5000 5.2000 0.4000

0.2491 5.0000 3.4000 1.6000

Each of these real numbers are an indivisible genes.

2.4. Population initialization

K cluster centers are accidentally selected from available list and insert in one chromosome. This trend repeats for all P chromosomes producing population. Indeed P is population size.

3.4. Fitness computation

Access of fitness contains two processes. In first process, clusters, base of centers is chromosomes, will produce. This means any point X_i specify to one of the clusters C_j with center of Z_j , if:

$\|X_i - Z_j\| < \|X_i - Z_p\|$, P=1,2,...,K and j≠P

After clustering is done, available cluster centers in chromosomes replace with average of any cluster's points. For cluster C_i , new center Z_i accesses in this way:

$Z_i^* = \frac{1}{n_i} \sum X_i$ i=1,2,...,K

For define this subject. We give an example. Assume that the first cluster center was (51.6 , 72.3). After clustering, there are 2 other points in this space:

(52.0 , 74.0), (50.0 , 70.0).So because this new center point of cluster 1 is identical with (10.0 + 74.0 + 7203) / 3 , (50.0 + 52.0 + 51.6) / 3 , this new center (51.2 , 72.1) replace with previous center (51.6 , 72.3) in relevant chromosome.

The clustering standard or □ is account in this way:

$\mu = \sum_{I=1}^{k} \mu_i$ $\mu_I = \sum \|X_j - Z_j\|$

Also fitness dependent accesses in this way:

$F = \frac{1}{\mu}$

4.4. Useable genetic operators

Crossover: Crossover is a probability trend that we change information among 2 chromosome called parent in order to produce 2 off springs. In this paper we apply one-point crossover with fixed crossover value P_c=0.8 .

Mutation: Each chromosome participates in mutation with a fixed probability. Supernatant decimal is apply for chromosomes. To mutation we use this way. A number of δ accidentally produces with monotonous distribution in span [0.1]. If the number of mutant gene is V , there is:

$V \pm 2*\delta*V$, $V \neq 0$

$V \pm 2*\delta$, $V = 0$

Positions of – and + happened with similar probability. For simplify in this project, mutation is done in the form of $V \pm \delta*V$.

5.4. Stopping criteria

In this work , calculation trend of fitness, crossover, mutation is done in amount of maximum repeated number.

5. The result of simulation and comparison of algorithms

Before draw a conclusion , we give some definition for kind of parameters. As previously said , the object was assessment of K-means clustering method and proposal trend of GA-clustering and comparison of their result. For this, we use some collections of different parameters. The example is Iris Data set. That shows different clusters of a flower and contains 4 feature of Petal's length, Petal's width , Sepal's length , Sepal's width in cm. The parameters are describe in 3 class: 1, 0.5, 0.

The number of parameters should be identical with 3. Other parameter collection is called Vowel that contains 871 record parameter. Each record parameter constitutes 3 index f_1, f_2, f_3. This parameter should be insert in6 class. Third parameter collection is called Crude oil that contains 56 record parameters and each record constitutes 5 index. This parameter should insert in 3 class.

GA-clustering implementation is done with these parameters:

Crossover rate (PC): 0.8

Mutation rate (PM): 0.001

Population size (pop-size): 100

Maximum iteration: 100

K-means clustering is done with these parameters:

Number of clusters or K: That there are different parameters in each kind.

Maximum iteration: 100

Observed results

The numbers in this table are clustering standard or μ; and the lower one can give better conclusion. This conclusion implement more and more and accept with different initiative population on different parameters collections.

Table 1-5: obtained μ from running two algorithms on Iris data set (k=3)

GA-Clustering	K-Means Clustering	Running times
30.1203	32.1988	1
29.6173	30.4915	2
29.9909	30.7811	3
29.6173	30.5039	4

Table 1-6: obtained μ from running two algorithms on Crude oil data set (k=3)

GA-Clustering	K-Means Clustering	Running times
278.965148	279.743216	1
278.965148	279.743216	2

278.965148	279.484810	3
278.965148	279.597091	4

Table 1-7: obtained μ from running two algorithms on Vowel data set (k=6)

GA-Clustering	K-Means Clustering	Running times
149346.489128	157460.164831	1
149406.751288	149394.803983	2
149346.152169	161094.118096	3
149355.823103	149373.097180	4

6. Conclusion

In this paper K-Means algorithm that is one of the popular clustering techniques has been surveyed and tried to apply one of the optimization method named genetic algorithm improve in unsupervised classification procedure. Genetic algorithms are population based methods that use from operators for processing of population chromosomes. In this research, we defined a representation of chromosome string and combine K-Means and GA together. Observing simulations in different runnings show that K-Means clustering based on Genetic algorithm improved clustering measurement μ better and more efficient rather than pure K-Means considerably.

References:

[1] Duda R. O., Hart P. E., and Stark D. G.,*Pattern classification*, second edition,John Wiley,2000.

[2] Melanie Mitchell, *An Introduction to Genetic Algorithms*, MIT Press, Cambridge, MA, 2002; first edition, 1996.

[3] Hu Shu-chiung, *Genetic algorithm-based clustering technique* , Ujjwal Maulik, Sanghamitra Bandyopadhyay , 2004.

[4] Jun Zhang, *Member, IEEE*, Henry Shu-Hung Chung, *Senior Member, IEEE*, and Wai-Lun Lo, *Member, IEEE, Clustering-based Adaptive Crossover And Mutation Probabilities for Genetic Algorithms, 2006.*

[5] Licheng Jiao, *Senior Member, IEEE*, Jing Liu, and Weicai Zhong, An Organizational Coevolutionary Algorithm for Classification, 2006.