UDC 519.25

PARALLEL GENETICS ALGORITHMS IN DATA CLUSTERING METHODS

Inoiatov R.Kh., Zhan Zhonglin
Lanzhou Jiaotong University

This article discusses the method of data clustering using parallel genetic algorithms. Selection of the optimal clustering result is based on a comparative analysis of several populations. The modified algorithm of clustering of «nearest neighbors» is taken as the basis of the proposed method. This choice provides an increase in the speed of selection of solutions, and also simplifies the structure of the chromosome in the genetic algorithm. Also for this method, it is not necessary to specify the initial number of clusters, which is an additional advantage.

Keywords: cluster, clustering, data processing, parallel GA, data mining.

Introduction. Clustering (or cluster analysis) is the task of splitting a set of objects into groups, called clusters, in such a way that objects in the same group (cluster) are more similar (for some attribute or criteria) to each other than to objects in other groups (clusters) [1]. At the same time, no preliminary assumptions about their structure are usually made.

Clustering is used at the initial stages of the study, when knowledge about data is small. If clusters are already found, it is possible to use other Data Mining methods to try establishing reasons for clustering or use clusters for classification and recognition. Most clustering methods are based on the analysis of the matrix of similarity measure (distance, conjugacy, correlation, etc.). If the criteria or metric is represented by distance, then the cluster is the group of points Ω such that the mean square of the intra-group distance to the center of the group is less than the mean square distance s^2 (variance) to the common center in the initial set of objects N.

$$d_\Omega^2 < s^2, \tag{1}$$

where $d_\Omega^2 = \frac{1}{N}\sum_{x_i \in \Omega}(x_i - x_\Omega)^2, x_\Omega = \frac{1}{N}\sum_{x_i \in \Omega} x_i$. (2)

The task of clustering has two problems: determining the optimal number of clusters and obtaining their centers and boundaries. The initial data for the task of clustering is the values of the parameters (attributes) of the research objects. Usually, determining the optimal number of clusters is the main researcher's task. As about boundaries of clusters, their definition can be automated by using various methods and algorithms.

To solve the problem of clustering by statistical methods, it is necessary to have conditional multidimensional density distributions of characteristics for each class. Then the task is to find a way of making the optimal decision about the belonging of the verified object to a particular class. Classical statistical methods not always can give an optimal clustering solution. And obtaining analytical model of conditional multidimensional distribution densities of the predicted parameter and attributes is a laborious process. It can be done by a separate independent task for each class of objects and for certain conditions for solving this task. But in real tasks it is not always possible to implement classical statistical methods [1]. For real tasks, even knowing the set of informative attributes (identification of which is a very laborious task), multidimensional conditional density distributions of attributes are not always available for study.

For solving clustering problems, it makes sense to use methods based on heuristic algorithms [2]. The concept of «heuristic algorithm» is that in this case the clustering algorithm does not strictly follow the theory, but is mostly based on the intuition and experience of the researcher. Such methods can give satisfactory results with limited initial information about the probabilistic characteristics of

© Inoiatov R.Kh., Zhan Zhonglin, 2018

the attributes and the predicted parameter. So, for using these methods for clustering, it is enough to have a set of attributes strongly correlated with the predicted parameter, and it is not necessary to know their conditional distribution densities.

Clustering methods based on the use of heuristic algorithms do not always lead to optimal solutions. However, in order to use them in practice, it is enough that the clustering error does not exceed the permissible value, and this can be achieved, for example, by selecting more informative attributes, using appropriate methods to improve the clustering.

A clustering method based on parallel genetic algorithm. Clustering can be considered as the task of constructing an optimal partitioning of objects into groups. In this case, the optimality can be defined as the requirement to maximize the density of clusters or to minimize the mean square distance between the cluster center and all its objects:

$$F_1 = \sum_{l=1}^{k} \sum_{i \in S_1} d^2 \left(X_i, X_l \right) \qquad (3)$$

Where l – is the cluster number (l = 1,2,...,k); X_l – is the center of the l-th cluster; X_i – a vector of values of variables for the i-th object, included in the l-th cluster; $d\left(X_i, X_l\right)$ – is the distance between the i-th object and the center of the l-th cluster.

To solve the optimization problem, we have to choose the appropriate method. In this task, the objective function (3) is multimodal (multiextremal), therefore it is preferable to choose the adaptive method of random search, which is the genetic algorithm (GA) [3].

The genetic algorithm is a heuristic search algorithm used to solve optimization and modeling problems by sequential selection, combination and variation of the desired parameters using mechanisms that resemble biological evolution.

The general scheme of GA in the context of the clustering problem is as follows:

1. In the beginning, there are randomly generated individuals and obtain a quality estimate for each solution, for example, by criterion.

2. According to their qualities individuals are chosen for creation of a new generation using evolutionary operators:

After choosing randomly a pair of individuals, crossover executes an exchange of the information within the pair with some probability. Creates a new solution based on recombination of existing ones.

Mutation is an operator for a slight change of one individual/several individuals in the population. It is random, so it is against staying in the local minimum. Creates a new solution based on a random slight change in one of the existing solutions.

Selection identifies the fittest individuals. The higher the fitness, the bigger the probability to become a parent in the next generation. Operator of choice ancestors, where is more prefer good solutions.

3. Repeat step 2 until an acceptable (optimal) result is obtained.

The main advantage of GAs within this task is that they have higher probability to find a best global solution. Crossover and mutation operators make it possible to obtain solutions that are different to the original ones – this way a global search is performed. Most popular clustering algorithms choose the initial solution, which then changes in the process of task solving.

The disadvantages of the classical GA are propensity to stagnate and effective work with tasks of small dimension.

The stagnation of the genetic algorithm is a state of the algorithm in which for a large number of generations there has been no change to best of fitness function of the population, but the current solution is very different from the global minimum. Adaptive genetic algorithm excludes periods of stagnation or reduces their duration to a minimum due to an increase diversity of the population.

Both of these disadvantages of the classical GA can be overcome by using parallel GA (PGA) [4, 5]. In PGA, there is always a selection-crossover-mutation cycle as in GA. A deme is one separated population (subpopulation) in many deme populations. Migration means an exchange rate of individuals between the demes. The exchange of genetic information between deme creates good conditions for providing variability. The algorithm more often finds a global optimum or number of calculations of the objective function is less than in the classical GA.

Since each individual in each deme is a solution of the clustering task, it is necessary to select the basic method of clustering and the structure of the chromosome, so that it will become possible to reduce the complexity of the algorithm for checking each individual for suitability by objective function [6]. For selection, two clustering methods were considered: the k-means algorithm and the modified nearest neighbor algorithm.

The k-means algorithm. The k-means clustering algorithm aims to partition n objects into k clusters in which each object belongs to the cluster with the nearest mean, algorithm constructs k clusters located at possibly large distances from each other. The choice of the number k can be based on the results of previous studies, theoretical considerations or intuition. Advantages: simplicity, speed of calculations, clarity of the algorithm. Disadvantages: the algorithm is too sensitive to emissions and noise, slow work on large databases, you must know the number of clusters.

The second method, based on joining the next point to the cluster in case if the distance between the new point and the previous point is less than a given threshold limit value. For each axis of attributes space, the following steps are performed:

1. Normalization of attributes space.

2. The density of the points location is calculated for each coordinate space of attributes.

3. The average distance r_i between two neighbor points is calculated on the selected coordinate axis, where i is the axis number:

$$r_i = \frac{1}{n-1} \sum_{t=2}^{n} d_t, \qquad (4)$$

Where $d_t = (x_t - x_{t-1})$; t – is the ordinal number of a point on the i axis as the coordinate value increases, t = *1...n*.

4. The degree of non-uniformity of distributions on each i-th axis (density dispersion) is calculated:

$$s_i^2 = \frac{1}{n}\sum_{k=2}^{n}\left(d_{ik} - r_i\right), \tag{5}$$

5. The coordinate axes are ranked according to the increase value of s_i^2.

6. Using procedure for distributing points on clusters for each i-th axis. We denote the clusters D_j, where j – is the cluster number, j = 1...n. Point x_1 is assigned to the cluster D_1. After this it is necessary to cycle $n-1$ following iterations:

6.1. Calculating distance $d_t = x_t - x_{t-1}$ between the points x_t and x_{t-1}, (t = 2...n, since the point x_1 is already in the cluster D_1).

6.2. If $d_t < \alpha_i r_i$, (α_i is the coefficient determining the cluster density threshold), then the point x_t belongs to the same cluster D_j as the point x_{t-1} , otherwise the cluster D_j is complete. The composition of this cluster is fixed, and the point x_t becomes the first point in the new cluster D_{j+1}, t increases by 1.

Then we perform step 6 for the i+1 axis. In this case, the points captured in the clusters defined on the previous axis are processed sequentially. Each of these clusters, after passing through step 6, can be divided into a number of smaller clusters.

After performing step 6 for the coordinate n, we get some number of the smallest clusters, which will be the result of clustering. This result and its estimation will depend on which values the threshold coefficients α_i taken.

Advantages of this method is that the preliminary specification of the number of clusters is not required if it is not necessary, and the clusters do not overlay.

The PGA strategy determined by the migration model (topology), which represent how the genetic information is exchanged between the demes [7]. Topology brings a new dimension to PGA, because we have got several demes instead of one. Demes exchange individuals among themselves and are not anymore controlled «globally». These demes evolve independently from each other over a period of time (called isolation period). After this, several individuals will move between the demes (migrate). The number of migrating individuals, the method of breeding individuals for migration, and the migration pattern determine how significant the genetic diversity and the exchange of information in the deme will be.

The choice of individuals for migration can be done:

– randomly (individuals are selected by random);

– on the basis of the objective function (the adapted individuals are chosen).

There is a wide variety of migration patterns of individuals between demes. For example, migration can take place:

– between all demes (topology of the complete graph);

– on the ring;

– between neighbor demes.

The remaining parameters of the algorithm – the number of demes, the number of individuals in the deme, the duration of the isolation period in each deme, are the same for all demes.

The PGA algorithm includes the following steps:

Step 1. Initialization of demes by random values of individuals.

Step 2. The implementation in each deme a given number of epochs of evolution. Allocation of population leaders. Copying coordinates of leaders in a separate array.

Step 3. Analyzing leaders for identity. It is possible that in the group of leaders there are points belonging to the same extremum.

To select such leaders, it is enough to compare the values of the objective function by the condition $\left|f^{L1} - f^{L2}\right| \le \varepsilon$, where ε is a small number. The coordinates of the leaders are also compared. If identical leaders are found, then one of them is replaced by an individual with a random value.

Step 4. For each leader:

4.1. The boundaries of the search for the extremum value are defined:

$$a' = x_i^* - A\left(b-a\right)/2, \tag{6}$$

$$b' = x_i^* + \frac{A\left(b-a\right)}{2}, \tag{7}$$

Where a' and b' – new search boundaries; a and b – initial search boundaries; x_i^* – coordinate of the found solution; A – size of the boundary of the extremum.

4.2. Start a multi-stage mutation and leader test cycle, to identify extreme values of the criterion in a given narrow range of working coordinate values. As a result, new values of coordinates and extremums are fixed in the leader array.

4.3. The array of leaders is sorted by ascending value of the criterion.

4.4. The first in the list of leaders pretends to replace the previous found value of the global extremum at this stage of the search.

Step 5. If the found solutions satisfied, the search is completed, otherwise go to Step 2.

The results of computational experiments. To compare the clustering results of the developed method and the k-means method there were used the mean square distance between the cluster center and all its objects, the minimum distance between clusters and the maximum intra-cluster distance [8]. The compared algorithms used the same metrics and the same objective function. For comparison by these criteria, the following calculations are required:

Minimum distance between clusters:

$$D_k = \max_l\left(\sum_{t,g\in L} d_{tg}^2\right), \tag{8}$$

Where d_{tg} – is the distance between the points «nearest neighbors» from clusters t and g.

Maximum intra-cluster distance:

$$D_k = \max_l\left(\sum_{i,j\in S_l} d_{ij}^2\right), \tag{9}$$

Where d_{ij} – is the distance between the end points of the cluster D_l.

For the tests, typical classification task with a dimension from 3 to 6 was chosen. An example of a test task is the classification of schools by four criteria: X1 – the number of buildings, X2 – the number of students in school, X3 – the number of students with high records, X4 – the area of buildings (Table 1).

The Euclidean metric was used to calculate the distances. The distances were standardized according to the formula:

$$z_{ij} = \frac{x_{ik} - x_k}{\sigma_k}, \qquad (10)$$

Where z_{ij} – standardized distance between the i-th and j-th objects, k – number of the criteria.

Table 1

Data for clustering

School number	X1	X2	X3	X4
1	4	2048	102	4101
2	1	962	48	1200
3	7	3913	196	7171
4	4	2664	133	4239
5	2	1702	85	2165
6	2	1630	82	2331
7	7	4039	202	7232
8	5	3685	184	5257
9	2	1612	81	2237
10	3	1854	93	3395
…	…	…	…	…
50	6	3678	184	6408

The diagram (Figure 1) presents the average estimates of the quality of clustering using the k-means method and the proposed method. As it can be seen, for all four criteria the proposed method has advantages. It should be added that the proposed method always ensures the separation of clusters in all coordinates, which is important in solving recognition tasks.

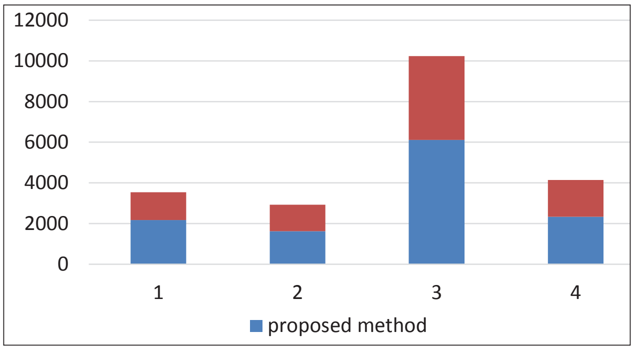

Fig. 1. Averaged assessments of the quality of clustering for test tasks

Conclusion. The advantage of PGA as a tool for solving the optimization problem is not only in the increase speed of the global extremum search process, but also in avoiding the stagnation of the global search process. By using several demes, it becomes possible to decompose the attribute space. As a base, a modified algorithm for forming clusters from «nearest neighbors» is used, which makes it possible to simplify the structure of the chromosome and to increase the speed of selection of the best solutions. The advantages of the proposed method on four criteria are demonstrated on test tasks. An additional advantage of the method is the guaranteed absence of overlay for all clusters and the absence of the need to know the number of clusters for algorithm.

Acknowledgment

This research is funded by the Natural Science Foundation of Gansu Province of China (1508RJZA076).

References:

1. Brian S., Landau S., Morven L., Stahl D. Cluster analysis 5th Edition. Wiley, 2011.
2. Alba E., Nebro A. J., Troya J. M. Heterogeneous Computing and Parallel Genetic Algorithms. Journal of Parallel and Distributed Computing, 2002, 62(9): 1362-1385.
3. Grajdeanu A. Parallel Models for Evolutionary Algorithms. George Mason University Press, 2003.
4. Akbari Z. A multilevel evolutionary algorithm for optimizing numerical functions. International Journal of Industrial Engineering Computations 2, 2011, 2(2) 419-430.
5. Thierens D. The Linkage Tree Genetic Algorithm. Parallel Problem Solving from Nature, PPSN XI. Springer Berlin Heidelberg, 2010, 15(11) 264-273.
6. Konfrst Z., Lazansky J. Extended Issues of PGAs based on one population. Neuro Fuzzy Technologies, 2002, 36(5) 71-78.
7. Sefrioui M., Periaux J. A Hierarchical Genetic Algorithm Using Multiple Models for Optimization. Parallel Problem Solving from Nature VI, 2000, 191(7) 879-888.
8. Skiena S. The Algorithm Design Manual. Springer Science+Business Media, 2010.

Іноятов Р.Х., Чжан Чжунлінь
Ланьчжоуський транспортний університет

ПАРАЛЕЛЬНІ ГЕНЕТИЧНІ АЛГОРИТМИ У МЕТОДАХ КЛАСТЕРИЗАЦІЇ ДАНИХ

Анотація

У даній статті розглядається метод кластеризації даних, реалізований на основі використання паралельних генетичних алгоритмів. Відбір оптимального результату кластеризації здійснюється на основі порівняльного аналізу кількох популяцій. За основу запропонованого методу взято модифікований алгоритм кластеризації «найближчих сусідів». Даний вибір забезпечує збільшення швидкості відбору рішень, і спрощення структури хромосоми в генетичному алгоритмі. Додатковою перевагою запропонованого методу є відсутність необхідності початкового визначення кількості кластерів.

Ключові слова: кластер, кластеризація, обробка даних, паралельні ГА, збір даних.

Иноятов Р.Х., Чжан Чжунлинь
Ланьчжоуский транспортный университет

ПАРАЛЛЕЛЬНЫЕ ГЕНЕТИЧЕСКИЕ АЛГОРИТМЫ В МЕТОДАХ КЛАСТЕРИЗАЦИИ ДАННЫХ

Аннотация
В данной статье рассматривается метод кластеризации данных, реализованный на основе использования параллельных генетических алгоритмов. Отбор оптимального результата кластеризации осуществляется на основе сравнительного анализа нескольких популяций. За основу предлагаемого метода взят модифицированный алгоритм кластеризации «ближайших соседей». Данный выбор обеспечивает увеличение скорости отбора решений, и упрощение структуры хромосомы в генетическом алгоритме. Дополнительным преимуществом предлагаемого метода является отсутствие необходимости изначального указания количества кластеров.

Ключевые слова: кластер, кластеризация, обработка данных, параллельные ГА, сбор данных.