Global Optimization Algorithms In Data Clustering Problems (August 2014)

Boris Zalmanovich Belashev¹*, Konstantin Anatolyevich Dolgii²

¹ Pushkinskaya st. 11, Petrozavodsk, Russia, 185910
² Universitetskaya st. 10, Petrozavodsk, Russia
*Corresponding author: belashev@krc.karelia.ru

Abstract:
Global optimization algorithms applied to solve the problems clear and fuzzy data clustering. With their help and using of different target functions the centers boundaries and forms of the clusters defined by modeling in MATLAB. It is shown that the centers of the clusters defined by algorithms patternsearch and K-means are the same. The advantages of global optimization algorithms are in the unity of approach, proximity solutions to an optimal solution, there is no requirements of continuity and differentiability of the objective function, taking into account the linear and non-linear constraints on the variables. The applications of the automatic clustering to processing tracks in high-energy physics, remote sensing of the Earth, geochemical data were discussed.

Keywords:
Optimization Algorithms; Data Clustering Methods; Functional; Centers; Boundaries; Forms of Clusters

1. INTRODUCTION

To perform data clustering, initial data are divided automatically into clusters without using a priori information. This procedure is employed to analyze and cluster data, to identify images and to extract information. Where information is structured and its main features are emphasized, clustering is a necessary stage largely responsible for a success in data processing in general [1, 2]. When clustering data, the number of clusters is estimated and their form, boundaries and position of the centres are assessed. A common technique is reducing the clustering problem to searching for the extremum of a function, e.g. a clustering variance minimum. In such cases, many known methods are often subjected to preliminary convergence: the local extremum they give may differ considerably from a global extremum [3]. The problem becomes especially significant when a data element belongs simultaneously to several clusters [4].

The goal of the present paper is to assess the potential use of global optimization algorithms [5] for cluster data analysis. The algorithms have two advantages: 1) the only near-optimum solution is obtained and 2) there are no requirements for the shape of a functional. Optimization algorithms are easy to use because they are realized in computer mathematics systems [6]. The modelling of demonstration examples
and problems using global optimization algorithms was performed in the Matlab system designed by Mathworks Inc.

2. DATA CLUSTERING METHODS

In agglomerative methods, all objects are originally considered as individual clusters that are then combined to form larger clusters. The combining process is clearly demonstrated by a hierarchic tree known as dendrogram [7]. In the main components method, the Karunen-Loew transformation [8] is used to pass to own matrix vectors for correlation of a system of arbitrary values and components in the form of a sum of data projections on these vectors that effectively diminish the total variance of arbitrary values. In a coordinate system with axes oriented along main components data are divided into clusters, and such a division can often be made meaningful, proceeding from the shape of the main components. The main components method is generalized in nonlinear manner in the form of self-organization maps of neuron models [9]. To refer objects to clusters in the mining clustering method, heuristic potential functions are used by selecting elements with the highest potential values as clustering elements [10]. In this method the number of clusters is not preset, as in agglomerative methods and in the main components method. If the number of clusters is known, the centers of clusters are obtained using the iterative method for \(-mean\) values [10]. The initial positions of the centers are selected arbitrarily. Following the “nearest neighbour” rule, objects belonging to each cluster are then selected, and its centre is calculated as an arithmetic mean or the gravity centre of its elements. The procedure is repeated until changes occur. Fuzzy clustering, based on the s-mean method, is used for overlapping clusters [11]. The centers of clusters are sought for using the functions of belonging of elements to the clusters and the number of clusters. The local extremes of fuzzy functions often yield the different values of the centers, depending on selection of initial points.

3. GLOBAL OPTIMIZATION ALGORITHMS

Genetic algorithms [12] implement C. Darwin’s natural selection idea – improving a species by transferring the best genes to descendants. The function of the environmental adaptability of individuals is determined on many individual chromosomes – a sequence of units and zeros that represent numbers in Grey’s code. The smaller the \(a\)-function value, the more environmentally adapted the individual. A genetic algorithm seeks for the global minimum of this function, beginning with an arbitrary set of individuals understood as a population. At each iteration, the individuals are combined to form pairs and produce descendants by crossing-over – exchange of chromosome tails and mutation - random value inversion in a random number bit. Individuals for a new population are selected with regard for the adaptability function of descendants and parents. The algorithm converges if a new population does not differ from the previous one. The advantage of the algorithms is that the functions need not be continuous and differentiable, they are not sensitive to being in local minima, multi-criterion optimization can be performed, convergence can be repeatedly accelerated in comparison with random search, and they can easily be realized. Acceleration of convergence is contributed to by codes that do not depend directly on target function arguments, the use of several points in the search space, and the determined and random solution search mechanisms. The disadvantages of the algorithms is difficult biological terminology, work with bit and pseudo-bit lines, complex coding of solution, inaccurate estimation of a global extremum, which is made more accurate by performing the algorithm several times and selecting an extremum value with the best adaptability function.
The pattern search algorithm, in which the pattern is the multiplicity of points in the form of the peaks of an n-dimensional cube that expands and contracts, depending on whether the pattern point has a value smaller than the current function value, is less consuming than genetic algorithms [13]. The minimum size of the pattern is the basis for search termination.

Biological or other special terms are used in global optimization algorithms. An ant algorithm searches for an optimum way in a column, imitates the behavior of the ants that blaze a trail from an anthill to a food source [15]. A particle swarm algorithm optimizes the function, supporting a solution population in the form of bees or particles that move in space in accordance with a simple law. The positions of the particles change when more advantageous positions are found [14]. An annealing algorithm is based on analogy with the state of crystal lattice upon cooling [15].

4. GLOBAL OPTIMIZATION IN CLUSTER ANALYSIS

If the number of clusters, \( m \), is known, then their centres can be determined from the condition of the global minimum of clustering variance. Let us express clustering variance \( D \) for \( n \) input objects with the coordinates \( u_i \) and \( v_i \), \( i = 1,2,\ldots,n \) and the centres of clusters with the coordinates \( x_k \) and \( y_k \), \( k = 1,2,\ldots,m \) by the relation (1)

\[
D = \sum_{i=1}^{n} \left[ (u_i - x_{r_i})^2 + (v_i - y_{r_i})^2 \right]
\]  

(1)

Here, \( x_{r_i}, y_{r_i} \) are the coordinates of the centre of the cluster with vector \( r_i \), that meet the condition:

\[
r_i = \arg\min_k \left((u_i - x_k)^2 + (v_i - y_k)^2\right)
\]  

(2)

The values of the variables that provide the global minimum of functional \( D \) are consistent with the coordinates of the centres of clusters.

Model data were processed using mining clustering methods, the K-mean method and the global optimization algorithm described above (Figure 1). The centres of clusters, obtained by the global optimization algorithm, differ from elements with the highest potential values (Figure 1(a)). The centres of clusters in the K-mean method coincide with those in the global optimization algorithm (Figure 1(b)), which is the result of efficiency: an unbiased estimate yields minimum variance.

The most adequate method for overlapping clusters is fuzzy clustering, which uses the functions of belonging of elements to clusters \( \mu_{ik} \) that take values from the interval \([0, 1]\). In this case, to evaluate the quality of partition, the criterion

\[
\sum_{k=1}^{m} \sum_{i=1}^{n} \left( \mu_{ik} \right) \left[ (u_i - x_k)^2 + (v_i - y_k)^2 \right] \to \min
\]  

(3)

is used, where \( \ell \) is exponential weight which is selected in the range \([0, \infty]\) and is normally equal to 2. To estimate the minimum (3), the s-mean algorithm with undetermined Lagrange multipliers is used [16]. To prevent untimely convergence, optimization (3) was performed using the pattern search algorithm [6, 13]. Processing of data by this algorithm has supported its ability to find a solution located near a global minimum.
Global Optimization Algorithms In Data Clustering Problems (August 2014)

Figure 1. Clustering of simulated data by the mining clustering method (a), the K-mean method and the global optimization algorithm (b).

We choose a straight line with coefficients A, B, C as the boundary between the two clusters whose elements have coordinates \( \{u_i, v_i\} \) in plane. These coefficients we determine from the condition:

\[
\rho = \left( \sum_{i=1}^{n} (\mu^l_{i1} + \mu^l_{i2}) \frac{Au_i + Bv_i + C}{\sqrt{A^2 + B^2}} \right)^2 \rightarrow \min \tag{4}
\]

Here \( \mu^l_{i1} \) and \( \mu^l_{i2} \) means the values of membership functions of i-th element to the first and second clusters, respectively, \( l \) is the exponential weight, selectable in the range \([1; \infty)\). For linearly separable clusters, \( \mu^l_{i1} \) and \( \mu^l_{i2} \) taking the values 0 or 1, \( l = 1 \), the condition \( (4) \) and a global optimization algorithm gives a unique solution that defines the boundary of the clusters (Figure 2).

Figure 2. Estimation of the boundary between clusters from condition \( (4) \).

Optimization algorithms can be used to visualize clusters in preset forms. Geometric figures or lines, such as circle or spiral arcs and segments of straight lines, are often used as the elements that form the framework of clusters. If data are clustered near a circle and a straight line, then clustering variance will take the form:

\[
D = \sum_{i=1}^{n} w_i \left( \sqrt{(u_i - x_0)^2 + (v_i - y_0)^2} - r_0 \right)^2 + \sum_{i=1}^{n} \eta_i \frac{(Au_i + Bv_i + C)^2}{(A^2 + B^2)} \tag{5}
\]
where \( x_0, y_0, r_0 \) are the coordinates of the centre and the circle radius, respectively, \( A, B \) and \( C \) are the coefficients of the line. Coefficients \( w_i, \eta_i \) take values equal to zero or unit, depending on condition:

\[
\begin{align*}
    w_i = 1, \eta_i = 0 & \quad \text{if} \quad \sqrt{(u_i - x_0)^2 + (v_i - y_0)^2} - r_0 \leq \frac{|Au_i + Bv_i + C|}{\sqrt{A^2 + B^2}} \\
    \eta_i = 1, w_i = 0 & \quad \text{if} \quad \sqrt{(u_i - x_0)^2 + (v_i - y_0)^2} - r_0 > \frac{|Au_i + Bv_i + C|}{\sqrt{A^2 + B^2}}
\end{align*}
\] (6)

On reaching the global variance minimum \( D \), the parameters of geometric lines are found to be most compatible with an input data set. Figure 3 shows the clustering of data that belong to a circle and three straight lines. If the lines that form the framework of clusters are preset in analytical form, their intersection points can be determined. Such an opportunity is of special interest for some applications.

Intersection elements are considered to belong to several clusters, and in description (6) instead of coefficients \( w_i, \eta_i \) we use the membership functions values \( \mu_{ik} \) and the fuzzy clustering algorithm

\[
D = \sum_{i=1}^{n} \sum_{k=1}^{m} (\mu_{ik})^l \varphi_{ki} \rightarrow \text{min} \] (7)

where \( \varphi_{ki} \) is the squared deviation of the \( i \)-th data element to \( k \)-th cluster. The results of fuzzy data approximation by a circle and a straight line are shown in Figure 4.

5. APPLICATION OF GLOBAL OPTIMIZATION TO CLUSTER ANALYSIS

Data clustering can be performed using global optimization algorithms in various fields. In high-energy physics, which is the study of the interaction of colliding particles, the goal at the initial stage
of data processing is to identify secondary particles and to determine their parameters from tracks [17]. Optimization algorithms are used to automatically relate the spots formed in detector material to straight line segments and circle and spiral arcs and to quantitatively estimate the parameters of these objects [18]. While performing these operations, special problems, such as search for interaction peak, are solved. The goal of automatic data clustering, performed while interpreting the results of the remote probing of the Earth and planets is to decipher the faults, the contacts of rock bodies, the boundaries of compositionally differing rocks and ring structures buried beneath the sedimentary cover. Such clustering is based on the minimization principle which means the maximum localization of prospects with a minimum number of the criteria used. Lineaments, large arcuate and ring structures, deep-structure blocks and transform and regional faults in the relief are normally sought for and their ore-controlling role is assessed. The goal of lineament analysis is to study the fracturing and deep structure of the crystalline basement and to assess tectonics-related risks. Ring structures, interesting in many respects, are formed upon domal uplift and during volcanic-tectonic processes and impact events. Attributed to uplift ring structures and the lineaments that cross-cut them are commercial mineral deposits, e.g. big gold, platinum and uranium deposits such as the Witwatersrand rift-related basin and the Muruntau deposit [19] or the Fennoscandian Shield’s biggest Kostomuksha iron deposit [20]. The ore fields controlled by local faults occur in the external belts of these structures. In geochemistry, clustering algorithms are used to reveal hidden relationships. Removing the apparently chaotic pattern of geochemical data, clustering contributes to their ordering and reveals homogeneous sets and anomalous values. Such an approach helps to better understand the mechanisms of processes and evolutionary changes and makes it possible to put forward and discuss hypotheses. Figure 5 shows the results of the clustering of chemical data for lamproite-lamprophyre-group rocks from the complex Kostomuksha deposit performed using the main component method [21]. Subdivision into three clusters is observed for factor F1 with positive coefficients in Mg, Ca, K, Al and Na and negative coefficients in Si and Fe, while vertical subdivision within one horizontally identified cluster is observed for factor F2 in which positive coefficients predominate in Mg, Ca, Al and Si and negative ones in Ti and Fe. The results obtained are considered to be due to the presence of phlogopite mica in diamondiferous lamproite group I-IV.
In information technologies, data clustering based on global optimization can be used in security systems, e.g. to reveal alien connections in network traffic, etc.

6. CONCLUSIONS

The application of global optimization algorithms to problems in cluster analysis has shown that they yield the only solution close to an optimum solution; the functions need not be continuous and differentiable; the restrictions imposed on variables are taken account of; and they converge rapidly. The algorithms, realized in computer mathematics systems and accessible to a broad circle of users, solve various optimization problems, determining the positions, forms and boundaries of clusters.

References


