Cluster Analysis: Some Theoretical Results and Practical Applications

A.K. Kerimov, R.A. Guliyev, R.I. Davudova, S.I. Abdulzade, J.M. Khalilova, U.Sh. Rzayeva

Azerbaijan State Economic University

adalat_kerim@mail.ru

Abstract. *The problem of automatic classification of objects described by quantitative features is considered. To solve this problem there is proposed the approach, consisting of a combination of two algorithms. Next, the solution of this problem is considered in terms of Boolean programming. There presented the analysis of the selection of combinations of informative features that define the object of cluster analysis, and a set of software algorithms of prognosis and enhance of the oil layer.*

Keywords

Automatic classification of the object, algorithm, sign, evolutionary algorithm, parameter, functional, informational content

1 Introduction

Cluster analysis problem is that on the basis of data about the set of objects is divided into clusters sets(subsets), so that each belongs to one and only one partition, and the subset of objects belonging to the same cluster were similar at the time as the objects belonging to different clusters are dissimilar.

Solution of this problem of cluster analysis is partitions that meet certain criteria of optimality. This criterion may be a functional expressing the desirability of different levels and groups of partitions, which is called *the objective function*.

Today there are many methods of cluster analysis. Let us discuss some of them.

2 Statement of the task

The problem of automatic classification of objects described by quantitative traits is under consideration.

Let there is a set of admissible, same type of complex objects

$$X = \{x_1, x_2, \ldots, x_m\},$$

which states are described by some set of numeric attributes

$$T = \|x_{ij}\|; x_{ij} \in M_j \subset R: i = \overline{1,m}; j = \overline{1,n}.$$

The task of automatic classification is according to calculate the values of predicates $P_q(x_i)$ "$x_i \in K_q$", $q = \overline{1,l}$, где $P_q(x_i) \in \{0,1,\Delta\}$ on the basis of the description of admissible objects, i.e.

$$P_q(x_i) = \begin{cases} 1, \ if \ x_i \in K_q \\ 0, \ if \ x_i \notin K_q \\ \Delta, \ if \ objects \ x_i \ do \ not \ belong \ klasses \ K_q, q = \overline{1,l} \end{cases}$$

3 Methods of solutions

In [5] to solve the problem there is presented the approach, consisting of a combination of two algorithms - the algorithm included in the family of algorithms for calculating estimates [14] and the evolutionary algorithm on the basis of Lamarck's principle [1]. The problem is solved in two stages. At the first stage for the initial value of the vector there are calculated binary proximity function of the object upon which an initial classification is constructed and the initial number of classes are determined. In the second phase to optimize the resulting classification there is performed the minimization of the functional characterizing the quality of the classification

$$J(\bar{u}) = \sum_{K_i \in K} \sum_{\substack{x_p, x_q \in K_i \\ p \neq q}} \bar{R}(p,q)$$

i.e.

$$J^* = J(\bar{u}^*) = \min_{\bar{u}(s) \in U} \sum_{K_j \in K(\bar{u})} \sum_{\substack{x_p, x_q \in K_j \\ p \neq q}} \bar{R}(p,q) \qquad (1)$$

where $\bar{R}(p,q)$ - distance between objects with numbers *p* and *q*, *U* – permissible area of three-dimensional vector of threshold estimates:

$$U = \{ \bar{u}(\varepsilon, \delta_s, \delta_k) \in [0,1] \} \qquad (2)$$

Here ε characterizes the proximity of the two objects on the basis of the generalized feature, δ_s - characterizes the general proximity of the two objects, a δ_k – the proximity of an object to a class.

In [13] to solve the problem (1) - (2) there were used methoods of Hooke-Jeeves and Rosenbrock and modifications, based on the nature of the tasks. In [2] an algorithm that optimizes the initial classification by transferring a set of objects from one class to another.

The self-organizing algorithm selects the priority scheme of reproduction, which reproduces the most successful descendants, surpassing their parents. This algorithm makes it possible to determine the effective number of parents. As mutations in the scheme there is used multipoint mutation and recombination of the interpolation. With the application of this algorithm the problem of classification of oil wells towards complications is solved. Many years of experience in operating of oil fields shows that the bottom-hole zone wells is the most vulnerable spot in the "layer-well". Production rates are highly dependent on the state of this zone. As the only link between the people and the object of their activities the bottom-hole zone is worst consequences of different technological influences. These effects occur near the wellbore, and in most wells, leading to various complications (salification, formation of sands, deposition of tar and wax, emulsion, formation damage by geological or technological reasons, corrosion, etc.)

In general, the given functional (1) is stair-stepped and has no property of unimodality and differentiability, which precludes the use of classical optimization methods.

In [3, 4] for solutions of this clustering problem we propose a new algorithmic approach, consisting of two heuristic algorithms - dynamic algorithm that is included in the family of algorithms for calculating estimates and self-organizing genetic algorithm, which selects the priority scheme of many parented reproduction and the effective number of parents. In [5] to solve the problem (1) - (2) there is proposed a new approach based on the Lamark's evolutionary principle.

Redistribution of objects between classes is an important issue of constructing of the optimal classification. To date, there is developed a number of decision rules, existing through a redistribution of objects between classes. In [11] there is found the criterion that provides a transition from "unfamiliar" class to "own" class on one object. In [6] there is considered generalization of changes in this criteria for case when μ objects simultaneously transferred from one class to another.

Let there is given a set *S* of *m* objects, *T*–*m*×*n*- dimensional attributive matrix. Each object is described by *n* features. Let it is known some preliminary classification. It is required to construct the optimal classification Φ by the selected criteria.

One of the most popular criteria (functional) is the sum of squared deviations, i.e.:

$$\Phi = \sum_{\nu=1}^{\tilde{l}} \Phi_\nu, \qquad \Phi_\nu = \sum_{x \in K_\nu} \left\| x - \bar{x}_\nu \right\|^2$$

where $\tilde{l}$ is number of pre-classes, and x are objects belonged to j-th class класcy, and $\overline{x}_j = \frac{1}{m_j}\sum_{\nu=1}^{m_j} x_\nu \quad , j \in I_{\tilde{l}} = \{1,2,...,\tilde{l}\}$ is the vector of averages for class K_j ; m_j - the number of objects belonging to the class K_j. Here $\|x\|$ denotes the Euclidean norm in R^n, that is $\|x\| = \sqrt{\sum_{i=1}^{n} x_i^2}$

In [6] there are concider two classes. Here we find conditions under which several objects simultaneously can be moved from one class to another class, in respect that the functional should be reduced.

Let simultaneously transferred μ - objects $x_p \in K_i$, $p = \overline{1,\mu}$ from class K_i to class K_j . Then there must be satisfied relation :

$$\Phi_i^* + \Phi_j^* < \Phi_i + \Phi_j$$

To move the μ objects from the class $K_i, x_\nu \in K_i$, $\nu = \overline{1,\mu}$ in the class K_j it is necessary and sufficient satisfaction of relation:

$$\sum_{q=1}^{\mu}\left\|x_q - \overline{x_j}\right\|^2 - \frac{\mu^2}{(m_j+\mu)^2}\left\|\overline{\mu} - \overline{x_j}\right\|^2 < \sum_{q=1}^{\mu}\left\|x_q - \overline{x_i}\right\|^2 - \frac{3\mu^2}{(m_j-\mu)^2}\left\|\overline{\mu} - \overline{x_i}\right\|^2$$

In [7] the above problem is considered in terms of Boolean programming.

We denote $|\ell_j|$ the number of objects in $j-$th class. “Costs” associated with the inclusion of the *i*-th object in the $j-$th group we denoted by c_{ij} .We introduce the variable x_{ij} as follows:

$$x_{ij} = \begin{cases} 1, \ if \ \ i-th \ object \ belongs \ to \ j-th \ class, \\ 0, \ in \ other \ \ case. \end{cases}$$

Then $|\ell_j| = \sum_{\ell=1}^{i_j} i, \qquad j = \overline{1,\ell}$.

If the number of classes ℓ is known before classification, then the matrixes of “cost” $C = \|C_{ij}\|$ and $X = \|X_{ij}\|$ will be of the order $m \times \ell$. In this paper the number ℓ is known, and after pre-classification of objects by the algorithms of calculating of estimating [14] it is defined the upper limit of the number of objects ℓ^*, $\ell \le \ell^*$. After the implementation of the model below, some preliminary classes may be empty, i.e. their number can be determined by defining of the content of the classes.

In the paper the matrix of “cost” C assumed to be known and intragroup sum of squares is taken as C_{ij}. Then the mathematical model of an optimal classification of objects will be in the following form:

$$\min\left\{\sum_{i=1}^{m}\sum_{j=1} C_{ij} X_{ij} \Big/ \sum_{i=1}^{m} X_{ij} = 1,\ X_{ij} = 0 \cup 1,\ i = \overline{1,m};\ j = \overline{1,\ell}\right\}.$$

It is known that the term “cluster analysis” includes a number of different classification algorithms, which grouped objects into clusters. Cluster analysis is one of the paradigms of technology Soft Coputing, which are constructed on the basis of the various systems of Artificial Intelligence. The concept of fuzzy logic is commonly used in two senses - the narrow and broad. In a narrow sense, fuzzy logic is a logic system, which is an extension of multi-valued logic. In the broad sense of the word, which is now dominated, fuzzy logic is equivalent to the theory of fuzzy sets, i.e., classes with

inaccurate, blurred boundaries. Thus, fuzzy logic, understood in a narrow sense, is a branch of fuzzy logic in the broad sense. Here, the question of belonging to the set is a matter of membership degrees.

Fuzzy logic in the broad sense (FLb) has the ability to expand the capabilities of classical logic in areas where classical logic can not provide satisfactory solutions [12]. There are some problems associated with natural language for which means of FLb can build a better model than it is possible in classical logic. This paper FLb serves in answering the questions posed in the theory of cluster analysis in the aspect of separability of objects into clusters within the framework of the rules. As part of this work two basic schemes of reasoning are described: an elementary deduction based on the modus ponens [8], as well as more complex scheme, which aims to establish a rigid or functional relationship between the objects and between objects and clusters.

To this day criterion for establishing of the completeness of the feature space does not exist, therefore, to establish the truth as more signs can be taken, then it is defined their optimal number by the synthesis of selected features. It is known that the information contains not only in the individual characteristics, it generally contains in the combination of features (informative combinations). One of the main problems of recognition task is to find those characteristics that lead to differences of objects from different classes.

For the construction of algorithms and programs in the problem tasks of cluster analysis, it is necessary to set rules for dealing with units of various levels of natural language. Generalized language units are defined by notion of Syntagm and syntagmatic descriptions may reflect one or another specific study. Syntagmatic turnover is more informative workload: they contain an additional message that accompanies the report, contained in the distributed parts of the proposal, and are characterized by the relative informative autonomy. In [9] description of the object by syntagm in terms of the scope and content of the object denoted by the word, is the basis of clustering. The algorithm of classification used different conditional clauses, which are the implication which is described in natural language. Next, the two-stage classification algorithm is used in the first phase of a preliminary set of clusters formed in the second stage, i.e. at the stage of improving the classification there is used the fuzzy approximation of the relationship between objects and classes.

One of the main problems in the oil industry is hard recoverable reserves. The effectiveness of development of low-productively deposits of oil depends on the quality and trouble-free operation of exploitation of the producing wells, the work of which is largely determined by the state of bottomhole zone. According to the current state of the bottomhole zone there is the greatest information on which it is possible to affect effectively the recovery and increase the productivity of oil wells. And when the old field starts to dwindle, the use of traditional methods of production often becomes ineffective.

The current stage of development of the oil and gas industry in terms of depletion of large deposits of oil production is characterized by the intensification and control selections, discovery and rapid commissioning of medium and small deposits, reduction of oil recovery, as well as the introduction of advanced techniques and science. Exploitation work is carried out generally not in individual wells, and on integral parts of fields and layers. And the problem of increasing oil wells is reduced to the problem of enhanced oil recovery. Therefore, in designated areas of the layer there are studied the structure and distribution of residual oil in detail, the dynamics of the main indicators of development, the types of complications, refined geological and physico-chemical conditions. On the basis of summarizing of information adopted geological and technical measures put into practice for a particular site.

The solution of this problem is reduced to the solution of the following sub-tasks in an environment of uncertainty considered [10]:

- selection of informative features that describe the state of oil wells;
- diagnostics state oil wells;
- forecasting state oil wells complications (saline, non-saline);
- choice of method or methods of enhanced oil recovery on the basis of two criteria - time and cost.

The software system was developed in DELPHI 7 to solve this problem.

4 Conclusion

Considered works in this article are the concise compilation of publications in the field of cluster analysis. The work object is achieved by various methods of automatic classification.

References

[1] Emelyanov V.V., Kureichik V.V., Kureichik V.M. Theory and practice of evolutionary modeling. - M., FIZMATLIT, 2003, 432 p.

[2] Kerimov A.K., Davudova R.I. An algorithm for constructing of the optimal classification//Proceedings of NANA.Vol.XXV, № 3, 2005. pp. 227-230.

[3] Kerimov A.K., Davudova R.I. Genetic algorithm in problems of automatic classification. / Proceedings of the Int. Conf. on Soft Computing and Measurements (SCM'2006). St. Petersburg, 26-29 June 2006 Volume 1. pp.250-252.

[4] Kerimov A.K., Davudova R.I. Implementation of self-organizing genetic algorithms in the optimal classification. / / Proceedings of NANA. Vol.XXV, № 3, 2008. pp.35-40.

[5] Kerimov A.K., Davudova R.I. An evolutionary algorithm for solving automatic classification//Journal "Artificial Intelligence and Decision Making." M.: 2009, № 4.

[6] Kerimov A.K. A new decision rule in the problem of automatic classification // Proceedings of the Academy of Sciences of Azerbaijan, Series of Physical-Technical and Mathematical Sciences, 2005, № 3, pp. 95-97.

[7] Kerimov A., Davudova R., Khalilova J., Huseynova H., Huseynova L. Construction of optimal cluster by evolutionary computation / XIII International Conference on Soft Computing and Measurements , 23-25 June, Sankt-Peterburg, 2012

[8] Kerimov A.K., Rzayeva U.Sh. The Problem of Functions' Fuzzy Interpolation Within Formal Theory// International Journal of Applied Mathematics and Statistics, 2011, v.27, №3, p.124-133

[9] Kerimov A.K., Rzayeva U.Sh. Clustering of objects using linguistic descriptions in the framework of the FL / / Artificial Intelligence and Decision Making, № 3, 2010, c. 95-100.

[10] Kuliev R.A. Diagnostic neural network system of states of the oil well functioning / First International Conference on Soft Computing Technologies in Economy, ICSCTE-2007. Baku, Azerbaijan. November 19-20, 2007. pp. 17-20.

[11] Tou J.T., González R.C. Pattern recognition principles. Addison-Wesley Pub. Co., 1974, 377 p.

[12] Zadeh L.A. The concept of a linguistic variable and its application to approximate reasoning I, II, III. Inf. Sci., 8, pp. 199-257, 301-357; 9, pp. 43-80, 1975

[13] Zenkin A.A., Zenkin A.I. The task of constructing of optimal classifications / / Collection of papers on mat. Cybernetics Computing Center of the USSR , Moscow, 1981, pp.20-33.

[14] Zhuravlev Y., Nikiforov V. Recognition algorithms based on the calculation of estimating // "Cybernetics", № 3, Kiev, 1971, pp.1-11.