Роза М.П. Методы решения задач автоматической группировки и задач размещения // Международный журнал информационных технологий и энергоэффективности. – 2023. – Т. 8 № 1(27) с. 21–23

Международный журнал информационных технологий и энергоэффективности

Сайт журнала:
http://www.openaccessscience.ru/index.php/ijcse/

УДК 004

МЕТОДЫ РЕШЕНИЯ ЗАДАЧ АВТОМАТИЧЕСКОЙ ГРУППИРОВКИ И ЗАДАЧ РАЗМЕЩЕНИЯ

Роза М.П.
Сибирский государственный университет науки и технологий имени академика М. Ф. Решетнева, Красноярск, Россия (660037, Красноярский край, город Красноярск, проспект имени газеты «Красноярский рабочий», д. 3) e-mail: mashenka-roza@mail.ru

В работе рассмотрены основные аспекты использования методов автоматической группировки в системах искусственного интеллекта; а также выявлена эффективность применения моделей оптимального размещения и автоматической группировки объектов; произведен анализ и классификация различных подходов для решения задач кластеризации.

Ключевые слова: автоматическая группировка, анализ данных, кластеризация, оптимизация.

METHODS FOR SOLVING AUTOMATIC GROUPING AND PLACEMENT PROBLEMS

Roza M.P.
Reshetnev Siberian State University of Science and Technology, Krasnoyarsk, Russia (660037, Krasnoyarsk Krai, Krasnoyarsk city, prospect named after the newspaper "Krasnoyarsk worker", 3) e-mail: mashenka-roza@mail.ru

The paper considers the main aspects of the use of automatic grouping methods in artificial intelligence systems; and also reveals the effectiveness of using models of optimal placement and automatic grouping of objects; analyzes and classifies various approaches to solving clustering problems.

Keywords: automatic grouping, data analysis, clustering, optimization.

Modern Russian researchers present the tasks of automatic grouping in the form of integer linear programming problems. Currently, there are a huge number of different effective methods for solving optimization problems. But it happens that difficulties arise during the solution of large tasks of automatic grouping, taking into account the fact that there is an instantaneous increase in the volume of data that is collected and processed in automated systems. Clustering, based on the established similarity relation of elements, establishes subsets (clusters) into which the input data is grouped. One of the simplest and most effective are methods and models based on minimizing the total distances between objects of the same group (cluster) or between cluster objects and its center.

Automatic grouping methods are used in many branches of science, including actively used in data mining systems [1]. Due to the use of models of optimal placement and automatic grouping of objects, the requirements for economic efficiency are increased. Automatic grouping methods can group objects by constructing models of the relationship of objects in a continuous space of characteristics. Such methods have the opportunity to be applicable with sufficiently large amounts of data. At the same time, tasks should be solved interactively with a limited working time with a large amount of input data.

The solution to the problem of automatic grouping is reduced to the development of a simple algorithm or an automated system as a whole that will be able to detect natural groupings in the data. There are two main types of methods in data analysis [5, 6]: research or descriptive, in which the researcher does not use predefined models or hypotheses, but tries to identify common characteristic properties of multidimensional data.

In the process of developing the direction of data analysis, many statistical methods have appeared: linear regression, discriminant analysis, variance analysis, multidimensional scaling, correlation analysis, factor analysis, cluster analysis [3, 4]. For pattern recognition tasks, the purpose of data analysis is to build a predictive model. Tasks in this formulation are also called learning. Learning tasks are divided into two classes: learning with a teacher (classification) – there is a training sample with data for which membership in a particular class is determined in advance; learning without a teacher is clustering in the complete absence of information about the affiliation of even a part of the data to a specific group.

The tasks of automatic grouping can be attributed to any of these classes: in both cases, the task of dividing a set of objects into homogeneous groups with similar characteristics should be solved. It should be noted that clustering is a more complex task.

The analysis and classification of the approaches proposed for solving the problems of automatic grouping of objects and data and cluster analysis is a difficult task due to their extreme diversity. Various approaches can use all kinds of similarity measures, various objective functions (minimizing the total distance between objects, minimizing the total distance to cluster centers, minimizing maximum distances in a cluster...). Groups can be defined as areas of high density in the space of features (characteristics) separated by areas of low density. And the algorithms themselves, based on this definition, are searching for connected high-density regions in the feature space, while different algorithms use different definitions of connectivity.

Algorithms using density reconstruction methods depend on the selected scale at which distances are measured, on the number of points falling into each other's neighborhood sufficient to determine such a cluster of points as a group, and on the maximum distance by which points in the group should be removed. The choice of the listed parameters is a difficult task, its solution determines the accuracy of the method and the adequacy of the automatic grouping model.

For example, such important tasks as image segmentation (in computer vision) are reduced to automatic grouping; grouping documents for their effective search, quick access and efficient use of memory during storage; splitting enterprise customers into groups in CRM systems for organizing effective marketing activities; tasks from the field of biology; tasks of recognizing printed and handwritten text.

Automatic grouping may also be required for natural classification (for example, organisms in living nature or inanimate objects), when structuring data, highlighting anomalies in data (identifying low-quality products in the production process), for data compression by replacing identical or very similar data objects with a single object that is their generalized (averaged) representation.

The idea of the work is to study the effectiveness of automatic grouping methods based on density for solving problems. The aim of the work is to assess the quality, accuracy and stability of the results when solving problems of automatic grouping and placement of several clustering algorithms when working with a large amount of input data.

To achieve this goal, it is necessary to perform the following tasks:

- To study the necessary literature on automatic grouping methods based on density, as well as on ways to assess the quality of clustering methods;
- Explore different assessment approaches;
- Study the problem statement;
- Develop an algorithm for the automatic grouping method based on density;
- To analyze the operation of algorithms on a test and real data set;
- Evaluate clustering algorithms on data sets;
- Analyze the obtained results of the selected algorithms;
- Draw conclusions about the work done.

The practical value of methods for solving problems of automatic grouping and placement problems is due to the wide range of their application both in cluster analysis or automatic grouping of data, and directly in practical problems of grouping physical objects or optimal placement in space.

For continuous placement tasks, algorithms have been developed to use simple and understandable metrics and distance measures. The arsenal of models used and the construction of universal methods for solving problems with various distance measures needs to be expanded. Among other things, there is an urgent need to improve methods for solving grouping problems with a large amount of input data.

Methods for solving automatic grouping problems can be divided into two large classes. The first class consists of methods that give a fairly accurate result, but at the same time require significant computational costs, and they are also applicable only for grouping a relatively small number of objects. The second class is specific methods that are aimed at quickly solving problems with a very large amount of data. These methods, in turn, often give a very inaccurate solution to the problem [2]. In addition – in some cases this is their most significant drawback - these methods give an unstable result, which strongly depends on the values of random variables, on the order of the data, and others.

It is important that when solving automatic grouping tasks, a method is used that meets the following criteria:

1. Solving problems in an acceptable time, allowing the construction of interactive automated systems for solving such problems.
2. The method should allow solving problems of automatic grouping of a large amount of data.
3. The method should allow solving problems using various grouping models based on the search for group centers.
4. For continuous tasks of automatic grouping, the method should be applicable with various metrics and distance measures that can be applied in practical tasks from various fields of knowledge.
5. The method should be combined with other local search methods used for specific tasks, as well as provide an opportunity to use various global search strategies.
6. The method should give stable results with multiple runs, while the accuracy of the results should not be inferior to other known methods with comparable counting time.
7. In addition to the fact that the method will solve problems of automatic grouping with a pre-known number of groups, so it still has to either give an estimate of the number of groups, or solve a series of problems with a different number of groups at once.

References

1. Cherezov D. S., Tyukachev N. A. Overview of the main methods of classification and clustering of data // Bulletin of Voronezh. state University. Ser. "System analysis and information technologies". 2009. Issue 2.
2. Ruban A.I. Methods of data analysis, Krasnoyarsk: CPI KSTU, 2004. 319p.
3. Sovigolovko E.V. Methods for assessing the quality of clear clustering // Computer tools in education, 2011 – pp 14 – 31.
4. Tabachnick, B.G. Using Multivariate Statistics, fifth ed. / B.G.Tabachnick, L.S. Fidell – Boston:Allyn and Bacon.– 2007.– p.980
5. Tukey J.W. Exploratory Data Analysis / J.W. Tukey. – Addison-Wesley. –1977. –P.688
6. Zhuravlev Yu. I., Ryazanov V. V., Senko O. V. "Recognition". Mathematical methods. Software system. Practical applications. – M.: Phasis, 2006. ISBN 5-7036-0108-8.