Anton USHAKOV, Zbigniew ŁUKASIK

APPLICATION OF CLUSTER ANALYSIS MECHANISMS FOR THE EXPLORATION OF CONTAINER TRANSPORTATIONS FUNCTIONING AT SELECTED RANGES OF THE TRANS-SIBERIAN RAILWAY

The article describes the basic methods and mechanisms of cluster analysis in relation to transport. In addition, an example of the analysis of individual polygons of the Trans-Siberian Railway is shown using a computer program that implements Kruskal and Prim methods.

INTRODUCTION

Recently, one of the most powerful toolkits that help to extract previously unknown knowledge from various, including large databases, is Data Mining Tools (DMT).

Data Mining Tools, also called Knowledge Discovery In Data, allow to significantly expand the range of practical management tasks that are solved using computers.

The discovery of new knowledge by means of data mining is carried out using a wide range of tools, among which an important place is occupied by cluster analysis.

The task of cluster analysis is to identify a natural local condensation of objects, each of which is described by a set of variables or characteristics. In the process of cluster analysis, the investigated set of objects represented by multidimensional data is divided into groups of objects similar in a certain sense, called clusters.

Cluster analysis is the basis of any intellectual activity and is a fundamental process in science. Any facts and phenomena must be ordered or grouped according to their similarity, i.e. are classified before general principles are developed that explain their behavior and mutual connection.

The result of cluster analysis is both the selection of the clusters themselves, and the determination of the belonging of each object to one of them. Often the results of the cluster analysis performed are the starting point for further data mining.

With the help of this further analysis, we are trying to establish: what is the revealed clustering and what is it caused by; who is a typical "representative" of each cluster; with the help of which "representatives" of clusters should solve various problem problems, etc.

In our case, the application of cluster analysis will allow us to reveal general patterns in the functioning of certain polygons of the Trans-Siberian Railway on some of their grounds, and thus to perform more efficient management, both by the safety systems of individual polygons, and by their functioning as a whole.

1. FORMALIZATION OF THE CLUSTERING PROBLEM

In the process of clustering, objects are grouped, to which anything, including observations and events, can be assigned.

The state of the object under study can be described using a vector of descriptors or a multidimensional set of attributes fixed on it:

$$X = \{x_1, x_2, \ldots, x_p\} \quad (1)$$

Then X_i is the result of measuring these attributes on the i-th object. Part of the signs can be quantitative and take any real values. The other part is of a qualitative nature and allows to order objects by the degree of manifestation of any quality (for example, a binary feature that reflects the presence or absence of this property).

Any multidimensional observation can be geometrically interpreted as a point in a p-dimensional space. It is natural to assume that the geometrical proximity of two or more points in this space means that these points belong to the same cluster.

To solve the problem of clustering algorithmically, it is necessary to quantify the concept of similarity and heterogeneity of objects. Then the objects X_i and X_j will be assigned to the same cluster, when the distance between these objects is sufficiently small, and to different ones - if it is large enough.

Thus, to determine the "similarity" of objects, it is necessary to introduce a measure of proximity or distance between objects.

There are different ways of calculating distances. The most commonly used is the Euclidean metric, which is related to the intuitive notion of distance.

$$\rho_E(X_i, X_j) = \sqrt{\sum_{k=1}^{p} (x_i^k - x_j^k)^2} \,. \quad (2)$$

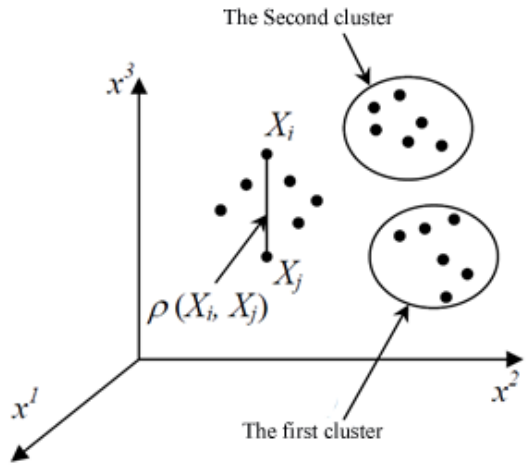


Fig.1. *Example of clustering*

Hemming distance is used as a measure of the difference in objects given by dichotomous (binary) traits. This measure is the number of mismatches of the values of the corresponding characteristics in the i-th and j-th objects under consideration:

$$\rho_X(X_i, X_j) = \sum_{k=1}^{p} \left| x_i^k - x_j^k \right|. \qquad (3)$$

There are other more abstract measures of intimacy. If the investigated characteristics are mixed (quantitative and qualitative), then the normalization of all the values of x_i^k of the quantitative characteristics x_k is necessary:

$$\frac{x_i^k}{\max\limits_i x_i^k}; \quad i = \overline{1,n}, \qquad (4)$$

which leads to a common Euclidean proximity measure.

When developing models and methods of clustering, it is usually assumed that objects within one cluster should be close to each other and far from objects that have entered into other clusters. The accuracy of clustering is determined by how close the objects of one cluster are and how far objects belonging to different clusters are deleted.

2. CLUSTERING ALGORITHM

Let the results of measurements of n objects be represented as a data matrix of size p × n, in which a set of rows represents objects, and a set of columns - signs.

$$\begin{matrix} X_1 \\ X_2 \\ \dots \\ X_n \end{matrix} \rightarrow \begin{pmatrix} x_1^1 & x_1^2 & \dots & x_1^p \\ x_2^1 & x_2^2 & \dots & x_2^p \\ \dots & \dots & \dots & \dots \\ x_n^1 & x_n^2 & \dots & x_n^p \end{pmatrix} \qquad (5)$$

Then the closeness between pairs of objects can be represented in the form of a symmetric distance matrix:

$$R = \begin{pmatrix} 0 & \rho_{12} & \dots & p_{1n} \\ \rho_{21} & 0 & \dots & \rho_{2n} \\ \dots & \dots & \dots & \dots \\ \rho_{n1} & \rho_{n2} & \dots & 0 \end{pmatrix} \qquad (6)$$

The general algorithm of cluster analysis, using a sub-algorithm for constructing a minimal spanning tree, contains the following main steps:

Step 0. [Initialization] Constructing the distance (proximity) matrix R from the measurement results of n objects represented by a data matrix of size p × n.

Step 1. [Construction of the minimal spanning tree] Using the matrix R, a minimal spanning tree T is constructed. To construct the minimal spanning tree, the Kruskal and Prim algorithms are used

Step 2. [Grouping objects into clusters] Vertices - the objects of the minimal spanning tree are grouped into clusters.

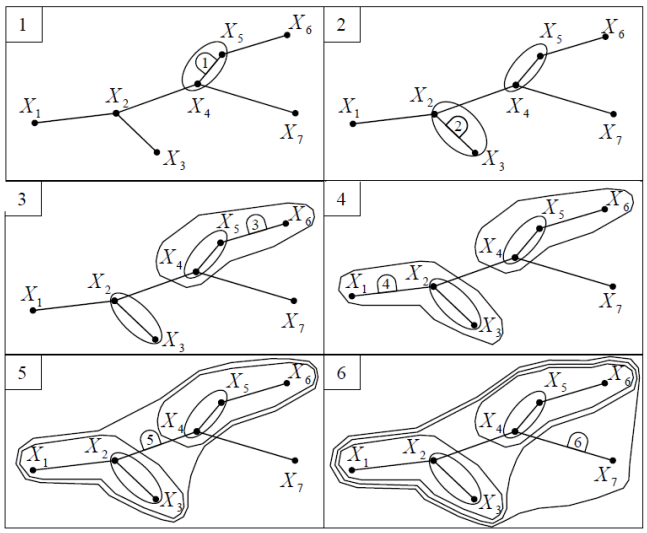


Fig.2. *Grouping Sequence of objects into clusters*

The order of combining objects into clusters can be specified using a parenthesis description. For the example under consideration, this bracketed entry has the following form:

$$\left(\left(\left(X_4, X_5\right), X_6\right), \left(\left(X_2, X_3\right), X_1\right)\right), X_7\right). \quad (7)$$

The most convenient and common way of describing the results of hierarchical clustering is the dendrogram:

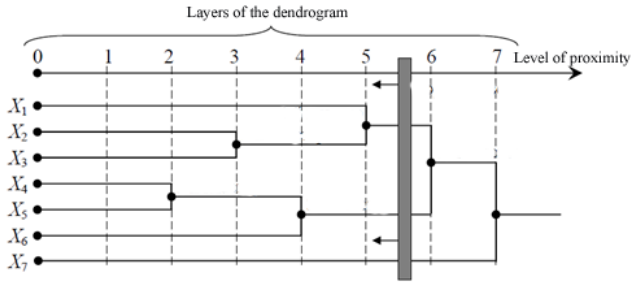


Fig.3. *Example of a dendrogram*

The dendrogram has a special tree structure consisting of layers of vertices, any of which represents one cluster. Each layer of vertices is characterized by its level of proximity. The location of an arbitrary cluster vertex relative to the layers of the dendrogram is determined by its level of proximity, which is measured by the weight of the last contracted edge in the formation of this cluster.

The formation of the dendrogram begins with a layer of zero level of proximity, in which each of the original objects is placed in a separate cluster. The lines connecting the vertices form clusters that are nested one in the other.

In general, the dendrogram reflects the nesting order of clusters, in which the number of clusters is successively reduced until a single cluster is formed that combines all the source objects.

The dendrogram cut, determined by its proximity threshold Δ, is used to perform cluster analysis on a given number of clusters. For this purpose, the proximity threshold Δ gradually decreases from the maximum possible value to zero. With such a decrease Δ, the dendrogram decomposes first into two clusters, then into three, etc., until the required number of clusters is met.

3. MECHANISMS FOR CONSTRUCTING A MINIMAL SPANNING TREE

The actions of the algorithms for constructing a minimal spanning tree T are considered on concrete examples of the distance matrix R.

Suppose we are given a symmetric distance matrix R, which can be associated with a weighted full-connected network G with n = 5 vertices and m = 10 edges:

$$R = \begin{vmatrix} 0 & 11 & 9 & 7 & 8 \\ 11 & 0 & 15 & 14 & 13 \\ 9 & 15 & 0 & 12 & 14 \\ 7 & 14 & 12 & 0 & 6 \\ 8 & 13 & 14 & 6 & 0 \end{vmatrix}, \quad (8)$$

Then the minimal spanning tree T of the network G is the cheapest subnet, i.e. a subnet of minimal weight that covers all vertices of the network G and contains no cycles. Obviously, such a subnetwork is a tree.

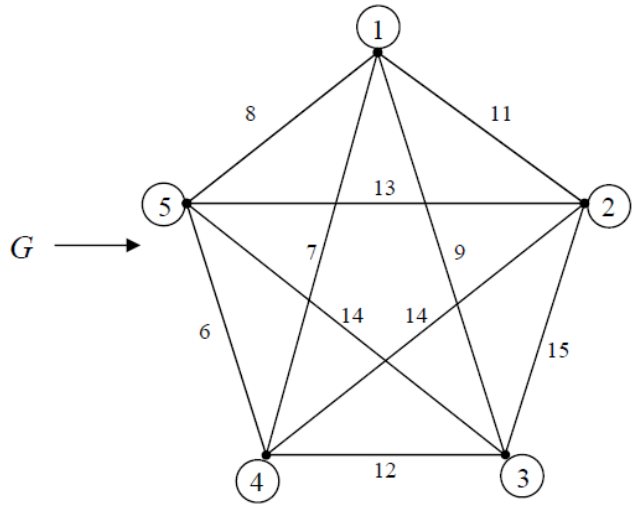


Fig.4. *Example of constructing a minimal spanning tree*

To construct a minimal spanning tree T in a weighted, connected and complete network G with n vertices and m edges, a number of algorithms can be used, among which the Kruskal and Prim algorithms are the most famous.

4. CHARACTERISTICS OF THE TRANS-SIBERIAN RAILWAY POLYGONS

It is well known that the Trans-Siberian railway is a railway through Eurasia, connecting Moscow with the largest East Siberian and Far Eastern industrial cities of Russia. The length of the highway is 9288.2 km, it is the longest railway in the world. The highest point of the way: Yablonovyy pereval (1019 m above sea level).

Historically, the Trans-Siberian is only the eastern part of the highway, from Miass (Chelyabinsk region) to Vladivostok. Its length is about 7 thousand km. This part was built from 1891 to 1916.

The result of the construction of the Trans-Siberian Railway was the opportunity created by 1905: for the first time in history, trains were used only on rails, without the use of ferry crossings, from the shores of the Atlantic Ocean (from Western Europe) to the shores of the Pacific Ocean (to Vladivostok).

Transsib connects the European part, the Urals, Siberia and the Far East of Russia, as well as Russian western, northern and southern ports and railroad exits to Europe, on the one hand, with Pacific ports and railway exits to Asia.

The starting point is the station Moscow-Passenger-Yaroslavl. The terminal station is Vladivostok. Throughput: 100 million tons of cargo per year.

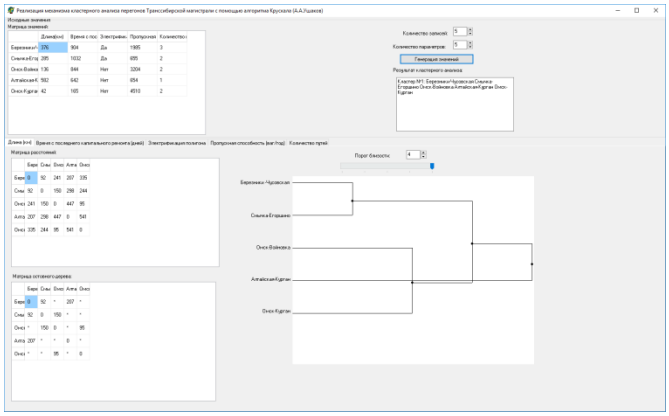
We will conduct a cluster analysis of the characteristics of the main railway stations throughout its historical route, specifically: Moscow-Passenger-Yaroslavl-Yaroslavl-Main-Danilov-Bui-Sharya-Kirov-Balezino-Vereshagino-Perm II -Kungur-Pervouralsk-Yekaterinburg-Passenger - Tyumen - Nazyvayevskaya - Omsk Passenger - Barabinsk - Novosibirsk-Main - Yurga I - Taiga - Anzherskaya - Mariinsk - Bogotol - Achinsk I - Krasnoyarsk-Passenger - Kansk-Yenisei - Ilanskaya - Taishet - Nizhneudinsk – Zima - Irkutsk Pas - Slyudyanka I - Ulan-Ude - Petrovsky Zavod - Chita II - Chernyshevsk-Zabaykalsky - Mogocha - Belogorsk - Birobidzhan I - I -Ruzhino Khabarovsk - Ussuriysk – Vladivostok.

For this we take the characteristics of Trans-Siberian Railway specific sections:

Name of railway section	Length (km)	Time since the last over-haul (days)	Electrification of railway section	Throughput (Railway carriages / year)	Number of paths
Berezniki-Chusovskaya	376	904	YES	1985	3
Smychka-Yegorshino	285	1032	YES	655	2
Omsk-Voynovka	136	844	YES	3204	2
Altayskaya-Kurgan	582	642	YES	654	1
Omsk-Kurgan	42	165	YES	4510	2
Lyangasovo-Nizhny Novgorod	390	640	YES	1645	1
Ljangasovo-Perm	376	846	YES	564	2
Perm-Agryz	109	1002	YES	1534	3
Agryz-Ekaterinburg	110	455	YES	3520	2
Smychka-Goro-blagodatskaya	143	1364	YES	1104	1
Voynovka-Bogdanovich	122	99	YES	1246	2
Orekhovo-Yudino	316	961	YES	854	1
Omsk-Altaiskaya	262	156	YES	3510	3
Bogdanovich-Kamensk-Uralsky	328	246	YES	946	3
Serov-Yegorshino	561	1062	YES	1581	2
Chelyabinsk-Kamensk-Uralsky	109	346	YES	796	3
Serov-Goroblagodatskaya	161	964	YES	779	1

In order to illustrate the possibility of using the Kruskal and Prim algorithms for cluster analysis of the operation of the Trans-Siberian Railway polygons, the author wrote a program that implements this algorithm using the Delphi 10 Seattle development environment.

The results of the clustering program in the screenshot below.



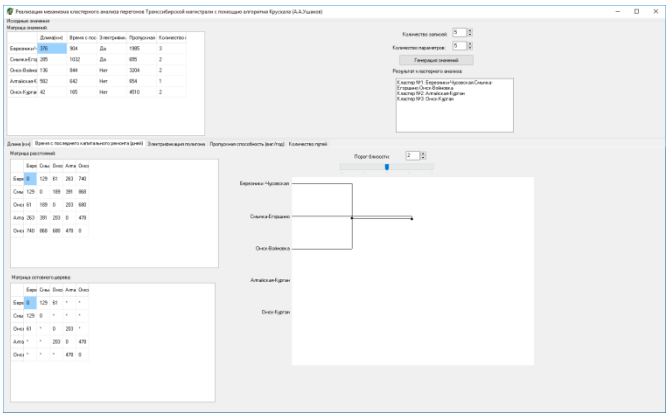


Fig.5. *The results of the clustering program*

CONCLUSION

In conclusion, it should be noted that cluster analysis is one of the effective methods that allows automating the grouping as separate polygons or stations of the Trans-Siberian Railway in order to provide visual models for management to improve service or upgrade existing railway lines.

BIBLIOGRAPHY

1. Joseph. B. Kruskal. On the Shortest Spanning Subtree of a Graph and the Traveling Salesman Problem. // Proc. AMS. 1956. Vol 7, No. 1. C. 48-50
2. R.C. Prim: Shortest connection networks and some generalizations. In: Bell System Technical Journal, 36 (1957), pp. 1389–1401
3. Thomas H. Cormen, Charles E. Leiserson, Ronald L. Rivest, and Clifford Stein. Introduction to Algorithms, Third Edition. MIT Press, 2009. ISBN 0-262-03384-4. Section 23.2: The algorithms of Kruskal and Prim, pp. 631–638
4. Kutyrkin A.V. Models and methods for developing large-scale subject domains for managing transport systems and production: Monograph. - Moscow: MIIT, 2004. - 148 p.
5. Zhambu M. Hierarchical cluster analysis and matching. - Moscow: Finance and statistics, 1998. - 342 p.

Zastosowanie mechanizmów analizy klastrów do poszukiwania transportów kontenerowych działających w wybranych zakresach kolei Transsyberyjskiej

Artykuł opisuje podstawowe metody i mechanizmy analizy skupień w odniesieniu do transportu. Ponadto przedstawiono przykład analizy poszczególnych wielokątów Kolei Transsyberyjskiej za pomocą programu komputerowego wdrażającego metody Kruskala i Prima.

Autorzy:
mgr. inż. **Anton Ushakov** – Russian University of Transport (MIIT), Russian Federation
prof. dr hab. inż. **Zbigniew Łukasik** – Uniwersytet Technologiczno-Humanistyczny w Radomiu, Wydział Transportu i Elektrotechniki

JEL: L92 **DOI**: 10.24136/atest.2018.243
Data zgłoszenia: 2018.05.29 **Data akceptacji**: 2018.06.15