T. A. Savchuk, Cand. Sc. (Eng), Assist. Prof.; S. F. Petryshyn

DETERMINATION OF EUCLIDEAN DISTANCE BETWEEN EMERGENCY CASES ON RAILROAD TRANSPORT DURING CLUSTER ANALYSIS

The paper analyzes the possibility of application of Data Mining technologies for analysis of emergency cases on railroad transport. There had been formalized the task of cluster analysis, revealed the main problems of distance ranging between emergency cases during such an analysis. There had been determined of usual, "weighted" and Euclidean distance squared case on railroad transport between the emergency cases.

Key words: *Euclidean distance, emergency case, railroad transport, cluster analysis,distance, degree of proximity,Euclidean distance square, «weighted» Euclidean distance.*

Introduction

An increase cargo transportation by railroad transport, taking into account the unhealthy and dangerous cargo, actualizes the problems of data analysis on emergency case which may occur during their transportation. The use of standard mathematical methods for such analysis allows the possibility for uncertain solutions, which using decision support systems, increase the risk of inexpedient or erroneous actions during abandonment job. Therefore, the reliable analysis objected on identification of such situations is of great importance during the development of decision support systems for their liquidation/decrease consequences [1].

Technology for data analysis, based on use of classical statistic approaches, suffer from a number of drawbacks when used for emergency cases analysis on rail road transport. These methods are based on the use of averaged indexes, which do not allow to determine the real state of such a situation. Methods of mathematical statistics appeared to be useful first of all, for checking the prior stated hypothesis and "rough" trial analysis, which is the base for in-line analytical data processing.

Apart from that, standard statistic methods do not consider non-typical observations, inadmissible during the analysis of emergency cases on railroad. Hence, some non-typical values may appear important for researches, characterizing the exceptional phenomena. The identification of these researches and their analysis with further consideration are necessary to realize the essence of the emergency case on railroad transport under research. As the modern researches show, such situations may become decisive in further behavior and development of an emergency case [1].

During the analysis of emergency satiation, which may appear during the transportation dangerous cargo, it is necessary to operate real values, organize the search for implicit regularity in data, independent framing hypothesis on independence of parameters and characteristics of such situations.

Survey of the existing methods for analysis of emergency cases on railroad transport

Great number of tasks and problems for analysis of emergency situations on railroad transport may be solved by Data Mining technologies depending on the character of the tasks to be solved (tasks for description and tasks for prognostication) as is shown on fig. 1. All the algorithms for data analysis are divided into supervised learning and unsupervised learning (fig. 1). In the first case the task for analysis is solved in some stages. First, using special algorithm, we build the model of data to be analyzed. Then the model considers study samplings until it starts working in a correct way. Unsupervised learning is used when there is no preliminary knowledge on data under analysis.

The main tasks of Data Mining are: classification, regression, search for association rules and clusterization (table 1) [2]. Let's consider their application on the example of analysis of emergency cases on railroad.

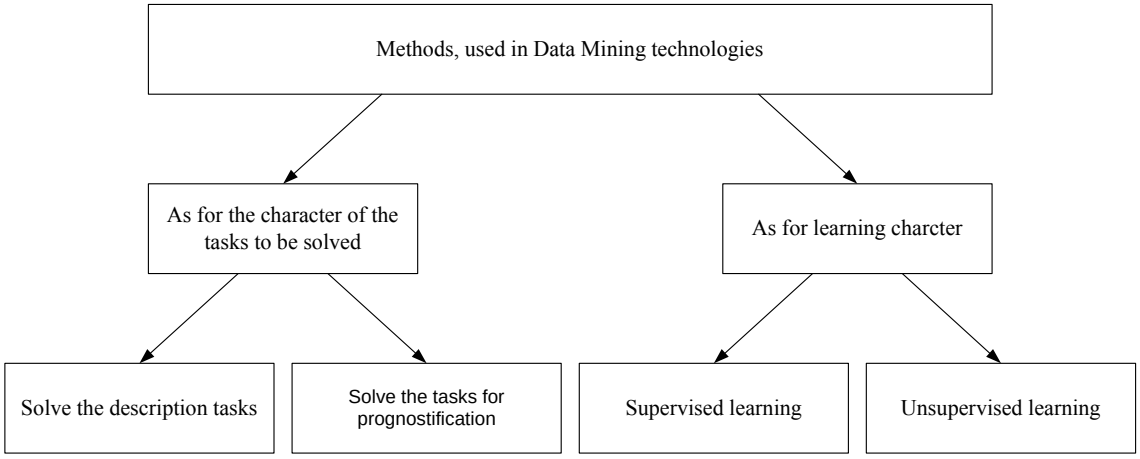

Fig. 1. Classification of methods, used in technology Data Mining to analyse emergency cases on railroad

The task of classification is reduced to the determination of class of emergency case on the railroad transport according to its characteristics. The set of classes of emergency cases may relay to, is known in advance. Classification in analysis the emergency case is very simple in use and is characterized by the availability of great number of efficient approaches to the solution of this task, but this is of no importance in issues under research. Apart from that, it should be noted that there are the following disadvantages of this classification: [2]:

1) capacity of learning sampling must be sufficient;

2) the learning selection must include emergency cases which represent all the classes, which is problematic during the analysis of such situations;

3) each class should have the sufficient set of emergency cases in learning sampling, which is difficult to receive during the analysis of emergency cases on railroad;

4) problem of overfitting, which means that the classification functions adapts well to the data, and if they have errors or outliers, the function interprets them as a part of internal data structure, which is unacceptable for analysis of emergency cases on railroad transport;

5) problem of underfitting, which means a big number of mistakes during verification of classifier, which is unacceptable for object sphere to be analyzed.

The task of regression like the task of classification allows to determine the value of its specific parameter following the known characteristics of emergency case on railroad transport. Unlike in classification, the value of this parameter is not the set of classes but the set of real numbers which is not important in the analysis of emergency cases on railroad transport.

The aim of search for associative rules is to determine dependences which are often repeated among the emergency cases. The found dependences are presented as rules and may be used for better understanding the nature of the analyzed data as well as for prognostication of appearance of specific events [2].

The advantages of search for associative rules is the fact, that they allow to find regularities among the emergency cases which is important for this subject area. Superficial perception of rules and simple interpretation of programming languages is not important during the analysis of such situations.

The disadvantage of search for associative rules is the fact that the rules which are found in the result of such an analysis, are not always useful, since there are three types of associative rules: useful, trivial, incomprehensible. Such an outcome is unacceptable for analysis of emergency cases on railroad.

The task of clusterization consists in the second for independent clusters in the set of data on the emergency cases on railroad transport to be analyzed. It allows to understand the data. Apart from that, grouping the homogeneous data allows to reduce their number to simplify further analysis [2].

Table 1

Characteristics of Data Mining tasks used for analysis of emergency cases on railroad transport

Name of task	Essence of task	Advantages of task	Disadvantages of task
Класифікація	Determination of classes of emergency case on railroad transport following its known characteristics	Simplicity in use, availability of big number of efficient approaches to the solution	Capacity of learning sampling must be big enough; learning sampling must include emergency cases which present all classes, each class must have capacitive set of emergency cases in learning sampling; problem of overfitting; problem of underfitting
Regression	Allows to determine the value of separate parameter following the known characteristics of emergency situation on railroad transport	Simplicity in use; availability of many approaches to the solution of the given task	Impossibility to solve the tasks of identification of emergency cases on railroad transport
Search for associative rules	Determination of dependences, which are often met between the emergency cases	Possibility to find specific regularities between emergency cases; superficial comprehensim of rules; constant interpretation of programming languages	Правила не завжди корисні, оскільки є три види асоціативних правил: корисні, тривіальні, незрозумілі
Clasterization	Search for independent clusters in set of data on the analyzed emergency cases on railroad transport	Iteration search for optimal result; possibility to use methods for cluster formation, choose the peculiarities of measures and proximity between two objects and cluster, object and cluster, two clusters; building the scientifically substantiated classification of multidimensional observations on the base of aggregate of the selected indexes and revealing of internal connections between the emergency case on railroad to be analyzed.	Determination on the input of number iterations during the search for solution

The advantages of clusterezation are the iteration search for optimal results which improves the probability in finding the optimal solution; possibility of use of methods for cluster formation and choose of features and measures of proximity between two objects, object and cluster two clusters, which is extremely important during the analysis of emergency cases on railroad; building of scientifically grounded classification of multidimensional observations on the base of aggregate of selected indexes and revealing internal relations between the emergency cases on railroad, to be analyzed.

The difficulty in realization of clusterization is resulted by the determination of number of iterations on the input during the search for solution, which is important the subject area to be analyzed, since it determines the exactness of predicted results in the operation of algorithm for identification of emergency situation which is subject to analysis and state of its development.

So, the results of the above analysis allow to make a conclusion on the expediency of use of clusterization for analysis of emergency cases on railroad transport. The data testify to the fact that clusterization is characterized by iteration search for optimal decision, possibility to choose features and measures of proximity between the two objects, object and cluster, two clusters, building of

scientifically background classification of multidimensional observations on the base of aggregate of selected indexes and reveal of internal connections between the emergency cases on railroad transport, which are to be analized.

Problems set-up

Cluster analysis is a way of grouping multivariate objects, which are the emergency cases on railroad transport. This analysis is based on presentation of results of separate emergency cases by points of separate geometric space with further singling the groups of these points out (clusters, taxons). Cluster analysis allows to single out compact, separated from each other groups of emergency cases, assuming the "natural" devision of set into the areas of clustering such situations, which allow to state the homogeneity of activities during the liquidation of emergency cases which belong to one cluster. Cluster analysis is used for analysis of emergency cases in the following situations [3]:

– data on emergency situation are presented as matrix of proximity or distances between the specific situations (1):

$$\Psi = \begin{Bmatrix} 0 & \psi_{12} & \dots & \psi_{1n} \\ \psi_{21} & 0 & \dots & \psi_{2n} \\ \dots & \dots & \dots & \dots \\ \psi_{n1} & \psi_{n2} & \dots & 0 \end{Bmatrix}, \quad (1)$$

where ψ_{ij} – distance between vector parameters $\psi(Y_i, Y_j)$, where Y_i and Y_j – emergency cases on railroad transport;

n – number of emergency cases on railroad transport, information on which must be processed (capacity of data base on emergency cases on railroad transport).

– data on emergency cases are presented as points in multivariate space (2):

$$Y = \{Y_1, Y_2, ..., Y_n\} = \begin{Bmatrix} y_{11} & y_{12} & \dots & y_{1m} \\ y_{21} & y_{22} & \dots & y_{2m} \\ \dots & \dots & \dots & \dots \\ y_{n1} & y_{n2} & \dots & y_{nm} \end{Bmatrix}, \quad (2)$$

where Y_i – specific emergency case on railroad transport;

y_{ij} – value of specific the j-$^{\text{th}}$ parameter of the i-$^{\text{th}}$ emergency case;

m – number of parameters of emergency cases, saved in data base.

It is possible to use two ways of data presentation to analyze the emergency cases. It requires to process the powerful data base, therefore, the second way in this case (2) is more expedient, since the proximity matrix during the increase in data base of emergency cases on railroad transport by one such case increases by $((l+1)^2 - l^2)$ elements where l - the initial number of emergency cases on railroad transport in data base. The drawback of the first way of data presentation is the fact, that

The addition of a new emergency case to the data base requires the calculation (using computer or not) of its proximity index or distance to each saved emergency situation in data base. And a person who makes a decision on liquidation of emergency case, is not aware of coordinates of each emergency case on railroad transport. As for the second way – the addition of the new emergency case to data base increases the number of the elements by m.

The advantage of the second way of data presentation is the possibility to determine the coordinates of each emergency case as the geometric object.

So the task of cluster analysis of emergency situation on railroad transport may be formulated as follows.

Let us assume that there is a set of emergency situations on railroad transport $Y = \{Y_i\}\ (i = \overline{1,n})$. Each of emergency cases has *m* characteristics. It is necessary to divide the statical *m*-dimensional range of change of values analyzed features of emergency situations into the intervals of grouping. That is, divide the set Y by $k(k \le n)$ clusters in a way, that the specific emergency Y_i situation would belong to only one cluster. The main condition in this case is the maximum of emergency similarity cases, which belong to one cluster and maximum difference in emergency cases from different clusters [3].

Distance and degree of proximity of emergency case on railroad transport

The difficulties in forming tasks of cluster analysis of emergency situations on railroad transport are associated with the determination of notion of their homogeneity [4] and weak data structuring.

In general case the homogeneity of two i - th and j - th emergency situation on railroad transport is determined by knowledge of rule for calculation of value ψ_{ij} which characterizes either the distance $a(Y_i, Y_j)$ between the objects Y_i and v from the researched set of emergency case on railroad transport $Y = \{Y_i\}\ (i = \overline{1,n}),$ or degree of proximity $\omega(Y_i, Y_j)$ between the same cases. If the set function $a(Y_i, Y_j)$ then the close in meaning to this metrics emergency cases are homogeneous, that is, they belong to one cluster. But it is necessary to compare $a(Y_i, Y_j)$ with specific threshold values, which are determined in each case. Thais approach is expedient to use for the determination of proximity measure $\omega(Y_i, Y_j)$ during the formation of homogeneous clusters on railroad transport.

The following requirements are to be observed:

– symmetry requirements ($\omega(Y_i, Y_j) = \omega(Y_j, Y_i)$);

– requirements to maximum similarity of energy cases ($\omega(Y_i, Y_i) = \max(\omega(Y_i, Y_j))$);

– requirements to distance correspondence between the emergency cases on railroad transport and proximity measure between them (if $a(Y_1, Y_2) \ge a(Y_2, Y_3)$ and $\omega(Y_1, Y_2) \le \omega(Y_2, Y_3)$).

Measuring distance between emergency situation

Distance between emergency cases Y_i and Y_j or metrix is an integral real function $a(Y_i, Y_j)$, if [5]:

– $a(Y_i, Y_j) \ge 0$ for all Y_i and Y_j from the set $Y = \{Y_i\}\ (i = \overline{1,n})$;

– $a(Y_i, Y_j) = 0$ if and only if $Y_i = Y_j$;

– $a(Y_i, Y_j) = a(Y_j, Y_i)$;

– $a(Y_i, Y_j) \le a(Y_i, Y_k) + a(Y_k, Y_j)$, where Y_i, Y_j and Y_k – any three emergency cases on rail road transport from the set $Y = \{Y_i\}\ (i = \overline{1,n})$.

During the cluster analysis of emergency cases there is a problem in measuring the distance between the separable cases. Main difficulties which appear, as the following [5]:

– ambiguity in choosing way of regulation;

– ambiguity in determining distance between the objects.

If we consider the results of the research of some emergency cases on railroad transport, then on fig. 2, following these results, we built the correlation field. Scale as for axis shall be selected arbitrarily. Fig. 2 (a) presents the singled out classes A, B, C the change of axis scale "Temperature" (fig. 2 (b)) are changed, and there will be formed classes A, which is identical to previous visualization, and B_1 which comprises classes B and C, which is unacceptable for the

analysis of emergency situations on railway road. It is impossible to determine the distance between the emergency cases in this case, since the factors are measured in different unites. It is necessary to standardize the indexes, that is to bring them to the dimensionless quality with an aim of correct measuring distance between emergency cases on railroad transport. Standardizing is a transition to the specific equal description of all features up to the introduction of new arbitrary unit, which allows to admit formal comparison of emergency case on railroad transport.

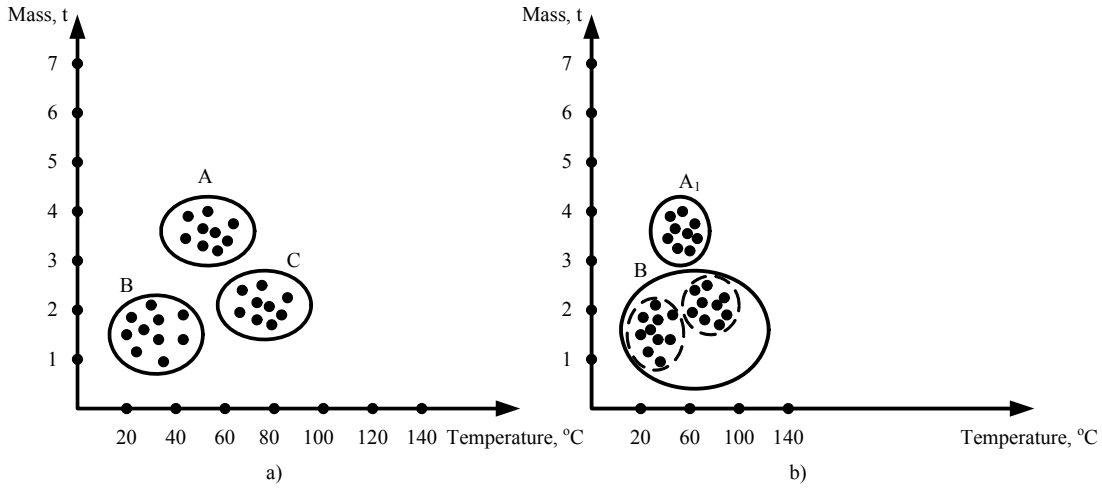

Рис. 2. Division of aggregate of emergency cases on railroad transport on clusters depending on scale of variables change.

Use of Euclidean distance for the analysis of emergency cases in railroad transport

Euclidean distance [6] is one of the used metrics in cluster analysis since it corresponds to intuitive conceptions of proximity and corresponds to classical statistic constructions by its quadratic form. This metrics is expedient to be used geometrically for uniting objects in aggregates which are typical for weak correlated sets.

The formula of general Euclidean distance looks like (3):

$$a_E(Y_i, Y_j) = \sqrt{(y_{i1} - y_{j1})^2 + (y_{i2} - y_{j2})^2 + ... + (y_{im} - y_{jm})^2}\ , \qquad (3)$$

where $a_E(Y_i, Y_j)$ – Euclidean distance between the two emergency cases on railroad transport Y_i и Y_j; y_{i1}, y_{i2}, …, y_{im} – vector of characteristics values, describing the j[th] emergency case on railroad transport; y_{j1}, y_{j2}, …, y_{jm} – vector of characteristics values, describing the j[th] emergency situation on railroad transport.

This metrics is expedient to apply in the following cases:

– values of parameters y_{i1}, y_{i2}, …, y_{im} are homogeneous in its physical respect, and if it is set, that all of them are equally important from the point of view of solving the task on referring the emergency case on railroad to definite cluster;

– space of features corresponds with the geometrical space of reality and the notion of proximity of emergency cases coincides with the notion of geometrical proximity in this space.

Consequently, the metrics may be accepted for the analysis of emergency cases on railroad transport to analyze characteristics, close in its physical meaning, which is unacceptable for this subject area, since it is necessary to analyze all the factors to receive reliable results.

The Euclidean distance squared is sometimes used to better distinguish between the distant objects [6] (4):

$$a_E(Y_i,Y_j)^2=(y_{i1}-y_{j1})^2+(y_{i2}-y_{j2})^2+...+(y_{im}-y_{jm})^2. \quad (4)$$

When it is necessary to determine the "weight" of each of characteristics λ_l of emergency case on railroad transport (for example, temperature in tank-car for transportation of highly inflammable substances and relative humidity) which will be proportional to the degree of its significance from the point of view of referring the emergency case to the definite cluster, it is expedient to use "weighted" Euclidean distance [4] (5):

$$a_{зE}(Y_i,Y_j)=\sqrt{\lambda_1\cdot(y_{i1}-y_{j1})^2+\lambda_2\cdot(y_{i2}-y_{j2})^2+...+\lambda_m\cdot(y_{im}-y_{jm})^2}, \quad (5)$$

where $a_{зE}(Y_i,Y_j)$ – "weighted" Euclidean distance between the two emergency cases on railroad transport Y_i and Y_j; λ_1, λ_2, …, λ_m ($0\le\lambda_l\le1(l=\overline{1,m})$) – vector of values of weight factor, corresponding to characteristics y_1, y_2, …, y_m of emergency situations on railroad transport; y_{i1}, y_{i2}, …, y_{im} – vector of values of characteristics, describing the j[th] emergency case on railroad transport; y_{j1}, y_{j2}, …, y_{jm} – vector of values of characteristics, describing the j[th] emergency case on railroad transport.

The learning sampling or experts experience is used for the determination of vector of values of weight factors λ_1, λ_2, …, λ_m. The attempts to determine the weight factors λ_1, λ_2, …, λ_m only on information which is in initial data give no result and may increase errors in the obtained result.

So, this metrix is acceptable for conducting cluster analysis of emergency cases on railroad transport, since it considers the importance of each characteristics of emergency situation on railroad transport, which improves the reliability of the result.

Conclusions

To analyze the emergency case on railroad it is expedient to use clusterization which is characterized by iteration search for optimal decision; possibility to choose features and proximity measures between two objects, object and cluster, two clusters; building the scientifically grounded classification of multivariate observations on the base of the selected indexes and revealing internal connections between the emergency cases on railroad.

Among the considered metrixes, the most expedient for the analysis of emergency cases on railroad transport is the weighted Euclidean distance, considering the value of each characteristics of emergency case on railroad transport which favors the obtaing of reliable result.

REFERENCES

1. Савчук Т. О. Використання ієрархічних методів кластеризації для аналізу надзвичайних ситуацій на залізничному транспорті / Т. О. Савчук, С. І. Петришин // Міжнародний науково-технічний журнал «Вимірювальна та обчислювальна техніка в технологічних процесах» (м. Хмельницький) – 2009. – №1 – С.193 – 198.

2. Барсегян А. А. Методы и модели анализа данных: OLAP и Data Mining. / А. А. Барсегян, М. С. Куприянов, В. В. Степаненко, И. И. Холод – СПб.: БХВ-Петербург, 2004. – 336 с.

3. Савчук Т. О. Порівняльний аналіз використання методів кластеризації для ідентифікації надзвичайних ситуацій на залізничному транспорті / Т. О. Савчук, С. І. Петришин // Наукові праці Донецького національного технічного університету. – Серія «Інформатика, кібернетика і обчислювальна техніка». – 2010. – Випуск 11(134). – С. 135 – 141.

4. Айвазян С. А. Прикладная статистика: Классификация и снижение размерности. / С. А. Айвазян, В. М. Бухштабер, И. С. Енюков. – М.: Финансы и статистика, 1989. – 607 с.

5. Мандель И. Д. Кластерный анализ. / И. Д. Мандель. – М.: Финансы и статистика, 1988. – 176с.

6. Дюран Б. Кластерный аналіз / Б. Дюран, П. Одел.; пер. с англ. – М.: Статистика, 1977. – 128 с.

Savchuk Tamara - Cand. Sc. (Eng), Assistant Professor with the Department of Computer Science.

Petryshyn Serhiy – Master in the Department of Computer Science.

Vinnytsia National Technical University.