Application of the Method of Data Mining to Solve the Problem of Establishing Causal Relationships

Tatyana Azhder

The article is devoted to solving the problem of identifying cause-and-effect relationships. *A method of constructing a domain model based on data mining, known as formal conceptual analysis, is considered. It allows to solve problems of structuring of the available information. The purpose of data analysis technology is the production of new knowledge, identifying relationships (links) in the data. This method is used in the article to solve the problem of diagnostics of technical systems.*

Keywords **: *cause-and-effect relationship, diagnostics, the formal conceptual analysis.***

I. INTRODUCTION

As you know, one of the main problems of prevention of technical accidents, as well as enormous losses from natural disasters, is the lack of the necessary level of provision of methods and means of measurement, testing, nondestructive testing, diagnostics and certification of products.

Technical diagnostics represents the theory, methods and means of detection and search of defects of objects of technical nature.

A defect is understood as any discrepancy between the properties of an object specified, required or expected. Detection of a defect is the establishment of the fact of its presence or absence in an object. The search for a defect consists in specifying with a certain accuracy its location in the object.

II. RESEARCH METHOD

The main purpose of technical diagnostics is to improve the reliability of objects at the stage of their production, operation and storage.

The reasons for the faulty and inoperable technical condition of the object can be detailed by specifying the relevant defects that violate the serviceability, operability or correct functioning and relating to one or more components of the object, or to the object as a whole.

To date, quite a lot of diagnostic systems have been developed that solve the problem of detecting defects of objects. But if the problem of diagnostics is put in another way, namely to consider it not only as a problem of detection of defects of objects, but also as a problem of identification of the reasons of emergence of these defects, i.e. the reasons of an inoperable condition of technical system, for its decision it is necessary to use system of formation of cause-and-effect relations. In this case, the external factors of the already existing situation can determine the actions and / or causes that affected the functional technical system and led to its faulty state.

The peculiarity of this problem is the need to structure the available information about the considered technical system. It should be noted that this information, at times, may be incomplete, as a result of which the search for the causes of the inoperable state of the object becomes difficult.

Data mining methods are used for automatic detection of empirical regularities and their use in solving problems of classification, pattern recognition, forecasting and diagnostics. The peculiarity of these methods is their orientation to problems for which the use of traditional statistical methods causes great difficulties [1-3]. For example, the tasks of analyzing a large amount of information, poorly conditioned tables (the number of features is comparable to the number of objects), etc.

The purpose of data mining is to produce new knowledge, identify relationships in the data. The methods of data analysis include the so-called formal conceptual analysis (FCA), introduced by Rudolf Ville. The method of constructing a domain model based on data mining is based on the theory of Birkhoff lattices and represents a recently formed logical-algebraic approach. FCA is applied to the structuring and formation of logical rules for establishing causal relationships.

This method is widely used in applied areas, including in solving the problem of identifying cause-and-effect relationships in the diagnosis of technical systems.

To describe the structure of the FCA, it is necessary to define its basic terms and properties [4-7].

A formal context is a triple (G, M, I) that consists of a set of objects G, a set of attributes M, and binary relations $I \subseteq G \times M$ between objects and attributes.

On the direct product G×M of two sets, there is a partial order if $(x_1, y_1) \le (x_2, y_2)$ if and only if $x_1 \le x_2$ in G and $y_1 \le y_2$ in M.

A lattice is a set L in which any two of its elements have an exact upper face, i.e. combining $x \vee y$, and an exact lower face, or intersection $x \wedge y$.

Revised Manuscript Received on November 05, 2019.

Tatyana Azhder. Assistant Professor at MIREA-Russian University of technology in the Department, which produces bachelors in computer science and masters in Information systems and technology.

Typically, a context looks like a table, with rows representing objects and columns representing attributes.

Let's imagine the context of failure of rotary equipment. In the rows we will write down the various causes of equipment failure (a set of objects G, indicated in the table by letters of the Russian alphabet), and in the columns-signs of these problems (a set of attributes M, indicated in the table by numbers). The signs will be descriptions of the spectra of the signal from the sensor measuring the vertical component of the acceleration of vibrations on the support of the equipment. The AC voltmeter measures the RMS, peak voltage and integral level of the measured signal.

Table- I: Context

Objects \ Attributes		1: At a frequency of 0..50Hz multiple jumps of SCR level up to 0.005 g	2: At a frequency of 0..50Hz one-time jumps of level of SKZ to 0,006 g	3: At a frequency of 0..50Hz single jumps of the level of scz more than 0,007 g	4: At a frequency of 0..50Hz single jumps of the level of scz more than 0,018 g	5: At a frequency of 100..150Hz single jumps of the SCR level up to 0.009 g	6: At a frequency of 100..150Hz single jumps of the SCR level more than 0.007 g	7: At a frequency of 150..200Hz multiple SKZ level jumps up to 0.004 g	8: At a frequency of 150..200Hz single jumps of the SCR level up to 0.006 g
Condition good	A		+			+	+	+	
Imbalance	B		+	+	+	+	+	+	+
The curvature of the shaft	D	+		+			+	+	+
Misalignment of the axes	F	+		+				+	

The table can be interpreted as follows. Each "+" character marks a pair that is an element of incident relation I.

For example, the binary relation (B, 4) means that in the case of an unbalance (object) on the spectrum of the sensor signal at a frequency of 0..50 Hz there is a jump in the value of the SCR level to 0.018 g (attribute). Thus, $(g,m) \in I$ means that "object g has property m".

The main concept in FCA is the formal concept. The concept (A, N) defines a pair of object $A \subseteq G$ and attribute $N \subseteq M$ that satisfy certain conditions.

A is called the extent, N is the intent of the concept, and the set of all the properties they possess is the content (intensional).

To determine the necessity and sufficiency of conditions for a formal concept, we present two operators, assuming $A \subseteq G$:

$$A^{'} = \{m \in M \mid \forall g \in A : (g,m) \in I\} \quad (1)$$

and accordingly for $N \subseteq M$

$$N^{'} = \{g \in G \mid \forall m \in N : (g,m) \in I\} \quad (2)$$

The above definitions mean that a set $A^{'}$ contains all the attributes that are common to all objects A, and a set $N^{'}$ is the set of all objects that have all the properties of a set N.

A pair (A, N) is a formal concept, if and only if

$$A^{'} = N \text{ and } A = N^{'}. \quad (3)$$

This property means the following: all objects of the concept contain all its attributes. This indicates that the misalignment of the axes when considering the spectrum of the sensor signal is characterized primarily by one-time jumps of the SCR level of more than 0.007 g at a frequency of 0..50Hz and multiple jumps of the SCR level up to 0.005 g at a frequency of 0..50Hz.

It follows from the definition of a formal concept that for all $A \subseteq G$ steam $(A^{''}, A^{'})$ there is a formal concept and for all $N \subseteq M$ steam $(N^{'}, N^{''})$ there is also a formal concept.

For formal concepts, the nature of the subconcept/superconcept relationship can be defined as follows:

$$(A_1, N_1) \leq (A_2, N_2) \Leftrightarrow A_1 \subseteq A_2, N_1 \subseteq N_2 \ . \quad (4)$$

This relation reveals a dualism between attributes and objects of concepts. A concept $C_1 = (A_1, N_1)$ is a sub-concept of a concept $C_2 = (A_2, N_2)$ if the set of its objects is a subset of the objects C_2. Thus, the set of all formal concepts form a so-called conceptual lattice.

If the context is given by a triple (G, M, I), then the infinum of such a lattice is formed by a set $\{\varnothing, M\}$, the supremum is formed by a set $\{G, \varnothing\}$.

Based on the dualism between objects and attributes, and the fact that data analysis and other applications of FCA interest us in the study of structures and relationships, it is necessary to present concepts that treat objects and attributes equally. This view is implemented in line charts.

A line diagram is a graphical representation of a conceptual lattice. It allows you to explore and interpret the relationship between concepts, objects and attributes, is an equivalent representation of the context. It contains exactly the same information as the relationship table, in which each node corresponds to a concept from a given context.

In the diagram, each object has properties assigned to a node and the properties of the nodes to which that node is connected by bottom-up arcs. At the same time, given the duality between objects and attributes, it can be argued that each property corresponds to the objects assigned to a given node, and those objects with whose nodes this node is connected by arcs from top to bottom.

On the basis of the theoretical descriptions of the FCA, the algorithm of its operation is derived.

Step 1. A record of the available data as a table in which the rows represent a set of G objects and the columns represent a set of M attributes.

Step 2. Establishing known relationships between objects and attributes (binary relationships $I \subseteq G \times M$).

Step 3. Define relationships between objects and attributes.

Step 4. Construction of a conceptual grid based on established concepts formed from a set of objects and a set of attributes

III. RESULT

Here is an example of constructing a lattice of concepts to analyze the cause of failure of rotary equipment. In table 1, the formal context K=(G, M, I) is given, where G is the set of causes of failure, M is their properties, and I is the binary relation between causes and properties.

When constructing a grid, the matching columns of the table (if any) can be interpreted as the presence of one or/and another feature. Therefore, if necessary, on the line chart, the second feature is enclosed in brackets. This situation can also indicate a linear relationship between columns.

Figure 1 shows the conceptual grid of the context "determining the cause of failure of rotary equipment".

A graph consists of nodes that represent concepts and edges that connect those nodes. Two nodes C_1 and C_2 are connected if and only if $C_1 \le C_2$ and there is no such concept C_3 that $C_1 \le C_3 \le C_2$.

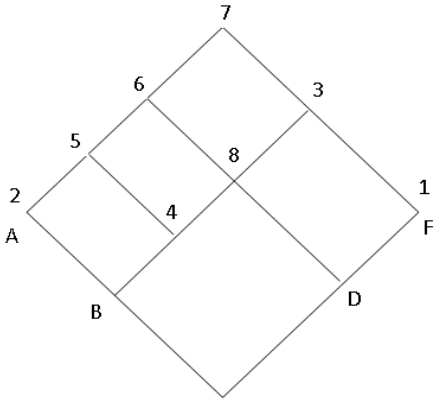

Fig. 1. The concept lattice of the context "determination of the cause of failure of rotary equipment"

Each object and attribute is entered into the graph only once. Attributes and objects are arranged along the faces of the graph as a kind of inheritance. Attributes are located along the faces to the top of the graph. Thus, the highest element of the graph (the upper edges of the context) corresponds to $\{\varnothing, M\}$. The base element of the graph (the lower edges of the context) corresponds to $\{G, \varnothing\}$. For clarity, attribute names are marked with numbers, and object names are marked with letters near the graph node.

Thus, with the help of a constructed graph, the relationships between objects and attributes become explicit and structured.

On the grid, you can trace all the properties that a particular problem of rotary equipment has: this is the set of all the properties that lie above the node marked with the name of the failure. Each lattice node corresponds to a concept.

Note that if for all context objects for which some property X is true, some property Y is also true, then the implication $X \to Y$ is true. In other words, if the implication $X \to Y$ is true for the context K=(G, M, I) and to any object $g \subseteq G$ every trait from the premise X applies, then to it also applies the trait from the conclusion of the implication Y, where $X \subseteq M$ and $Y \subseteq M$.

IV. CONCLUSION

Having considered the basic properties and working principle of formal conceptual analysis, it can be concluded that this method makes the connections between concepts explicit. The method is the best suited for the formation and formalization of concepts in poorly structured problems. Deciphering the conceptual lattice obtained as a result of the algorithm does not require additional knowledge, because it is quite simple and intuitive. Formal conceptual analysis can be used to solve the problem of identifying cause-and-effect relationships, as well as diagnostics of technical systems in the presence of incomplete information. This method can be used in solving problems of classification and structuring of knowledge, for example, in object-oriented design of mathematical software.

REFERENCES

1. A. Sigov, E. Andrianova, D. Zhukov, S. Zykov, I.E. Tarasov, Quantum informatics: overview of the main achievements. *Russian Technological Journal*. 2019;7(1):5-37. (In Russ.)
2. V. Rafalovich, Data mining, or data Mining for the employed. Practical course. Ed. Litiganti-Treyd, 2014.
3. A.A. Pastushkov, V.K.Batovrin, Selection of solutions for designing open systems based on analysis of variants with random weights. *Russian Technological Journal*. 2018;6(4):78-88. (In Russ.)
4. G. Birkhoff, T. Barty, Modern applied algebra, M., LAN, 2005.
5. B. Ganter, G. Stumme, R. Wille, Eds., Formal Concept Analysis: Foundations and Applications, Lecture Notes in Artificial Intelligence, State-of-the Art Series (2005), vol. 3626, pp. 196-225/
6. B. Ganter, G. Stumme, Creation and Merging Ontology Top-levels, Proc. 13th Int. Conf. on Conceptual Structures, ICCS'06, P. Hitzler, F. Sharfe, Eds., Lecture Notes in Artificial Intelligence, (2006).
7. J. McLennan, CH. Tang, B. Krivat Microsoft SQL Server 2008. Data Mining-data mining. –SPb.: BHV-Petersburg, 2009.

AUTHOR PROFILE

Tatyana Azhder. I am a candidate of technical Sciences. Work assistant Professor at MIREA-Russian University of technology in the Department, which produces bachelors in computer science and masters in Information systems and technology. My research interests are decision-making systems, data mining, artificial intelligence, neural networks.