RESEARCH PAPERS
FACULTY OF MATERIALS SCIENCE AND TECHNOLOGY IN TRNAVA
SLOVAK UNIVERSITY OF TECHNOLOGY IN BRATISLAVA

2016 Volume 24, Number 39

PREPARATION AND CLUSTER ANALYSIS OF DATA FROM THE INDUSTRIAL PRODUCTION PROCESS FOR FAILURE PREDICTION

Martin NÉMETH, German MICHAĽČONOK

SLOVAK UNIVERSITY OF TECHNOLOGY IN BRATISLAVA,
FACULTY OF MATERIALS SCIENCE AND TECHNOLOGY IN TRNAVA,
INSTITUTE OF APPLIED INFORMATICS, AUTOMATION AND MECHATRONICS,
ULICA JÁNA BOTTU 2781/25, 917 24 TRNAVA, SLOVAK REPUBLIC
e-mail: martin.nemeth@stuba.sk, german.michalconok@stuba.sk

Abstract

This article is devoted to the initial phase of data analysis of failure data from process control systems. Failure data can be used for example to detect weak spots in a production process, but also for failure prediction. To achieve these goals data mining techniques can be used. In this article, we propose a method to prepare and transform failure data from process control systems for application of data mining algorithms, especially cluster analysis.

Key words

Predictive maintenance, data mining, clustering, distance matrix, production process

INTRODUCTION

The term Big data is very topical in industry. Most industrial companies collect huge amounts of data from different processes in production, which have significant information value. One of the topical issues in industrial big data analysis is failure analysis and prediction. To discover any of the possible relationships between emerging failures in the production process, it is appropriate to use data mining methods. According to (1) data mining is a set of methods used to discover relationships in data in large databases. It overlaps to some degree with fields like artificial intelligence, machine learning, pattern recognition and data visualization. There are however many other definitions which largely depend on the use of data mining. According to Fayyad, data mining is the nontrivial process of identifying valid, novel, potentially useful, and ultimately understandable patterns in data (6). Zekulin defines data mining as the process of extracting previously unknown, comprehensible, and actionable information from large databases and using it to make crucial business decisions (7). Parsaye's definition says, that data mining is a decision support process, where we look at large databases for unknown and unexpected patterns of information (7). Data mining is a set of various methods to achieve an established goal. This paper deals with clustering, which is one of the methods of data mining. Clustering algorithms attempt to automatically partition the data into

a set of regions, called clusters, in which all data are similar to each other. This similarity is most often defined by the Euclidean distances between data. Clustering algorithms are often used to realize the hidden structure, or existing patterns in data. For the purpose of finding relationships in failure data, we have chosen to use the hierarchical clustering method. Hierarchical clustering is thought to produce better quality clusters. This method is known as a set of flat clustering methods organized in a tree structure. Hierarchical clustering is based on recursive partitioning of the data in a top-down or bottom-up structure, where the bottom-up approach tends to be more accurate, but has higher computational requirements. The input to a clustering algorithm is proximity (or distance) matrix. One of the aims of this article is to propose a method to compute a distance matrix for failure data from the industrial production process.

DATA PREPARATION

When dealing with data from real life, data preprocessing of some kind will be necessary. Most common problems with real data could be for example missing values, duplicate records, or contradicting records. Before applying any of the data mining methods for knowledge discovery from the data, it is necessary to check the data for these errors and if possible to correct them.

The data for the research described in this paper was acquired from the automotive industry. The data was exported in an .xls file, where each parameter is stored in a separate cell. Each failure has various parameters like localization, start time and date, end time and date, computed duration of the failure and a short predefined description. To find possible missing values or duplicate records a small application was written in C# programming language in the Visual studio 2015 environment using proper COM objects to work with .xls files. The search resulted in 0 missing values and 10 duplicate records. These records were deleted from the data set.

Real data also suffers from noise. This noise is understood as data records that have almost nothing in common with the rest of the data set. These records are statistically insignificant and thus they can be removed or ignored. Our research is aimed at finding relationships like sequential patterns in failure data. Because of this, we have ignored failure types with repetition number, across the whole data set, less than 10.

The next step in the data preparation for the application of cluster analysis was the data normalization. As mentioned before, the input to the clustering algorithm is a distance matrix. To compute the distance matrix all necessary parameters should be in a numeric format. For the purpose of normalization, we have assigned a class number to each failure type, predefined description of the failure and failure localization. Parameters start date and time, and end date and time were in dd/mm/yyy hh:mm:ss format. For these parameters, we have converted them into integer values which represent seconds. The failure duration parameter was already in the correct format, so it was not necessary to normalize it.

PROPOSED METHOD FOR DISTANCE MATRIX CALCULATION

After the data preparation phase, it is possible to transform data into a form to which a clustering algorithm can be applied. To use cluster analysis with the hierarchical clustering method, it is necessary to compute a distance matrix from a given data set. In our case, the distance matrix captures distances between each failure type computed as a Euclidean distance value. This value is understood as the square root of the sums of the squares of the differences between the coordinates of the points in each dimension. So the Euclidean distance between two random points $[x_1, x_2, \ldots, x_d]$ and $[y_1, y_2, \ldots, y_d]$ is computed as follows (2):

$$\sqrt{\sum_{i=1}^{d}(x_i - y_i)^2}\,.$$

A Euclidean distance matrix in $R_+^{N\times N}$ is an exhaustive table of distance-square d_{ij} between points taken by pair from a list of N points $\{x_\ell, \ell=1 \ldots N\}$ in R^n. Each point is labelled ordinally, hence the row or column index of a Euclidean distance matrix, i or j =1 ... N, individually addresses all the points in the list (3).

The distance matrix from production failure data was also computed using Euclidean distance between failure types. First of all, it was necessary to describe each failure type by chosen properties, so each failure type can be represented as a point in the Euclidean space. From the set of parameters, start date and time and duration of the failure were chosen. With the use of the first parameter "start date and time", the time distances between the same failure types emerging were calculated. These distances were then calculated for each failure type. Next, the average time between emerging failures of the same type were calculated also for each failure type. Subsequently, the average failure duration time was calculated for each failure type. After calculating these values, each failure type was described by two parameters (average time between the same failure type emerging, average duration of the failure type) which could be also understood as coordinates of a point in the Euclidean space. Then the failure type itself could be considered as a point in the Euclidean space. From these computed values, it was further possible to calculate the distance matrix itself with the formula for calculating the Euclidean distance.

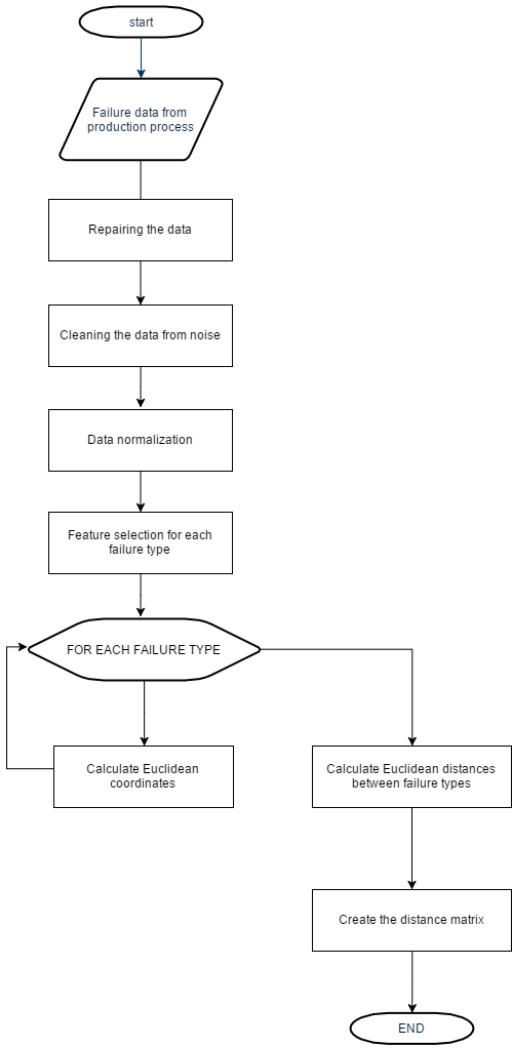

Fig. 1 *Flowchart diagram of data preparation process and distance matrix creation*

FINDING RELATIONSHIPS IN FAILURE DATA USING HIERARCHICAL CLUSTERING

As mentioned before in this paper, the calculated distance matrix serves as an input to clustering algorithms. In this case, when the matrix was used as an input to the hierarchical clustering algorithm. The hierarchical method can be subdivided as following (4, 5):

- Agglomerative hierarchical clustering: A bottom-up approach whereby each object initially represents a cluster of its own, then similar clusters are iteratively merged until the desired cluster structure is obtained. This algorithm for N samples begins with N clusters and each cluster contains a single sample. Afterwards two clusters with the closest similarity will merge until the number of clusters becomes one or as specified by the user. The criteria used in this algorithm are min distance, max distance, average distance, and center distance.

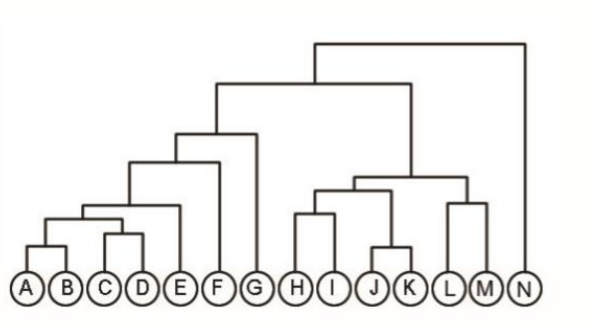

Fig. 2 *The dendrogram of an agglomerative hierarchical clustering algorithm*

- Divisive hierarchical clustering: A top-down approach where all objects initially belong to a single root cluster and iteratively partitions existing clusters into sub-clusters.

For analyzing the failure data, an agglomerative hierarchical clustering approach was used. The cluster analysis was performed in Statistica 13 software. This software consists of a fully integrated line of analytic, graphics and data management solutions. The data mining module of this software solution provides multiple methods and algorithms including hierarchical clustering. The output of this method is a set of clusters organized in a tree structure. The following table shows cluster membership of each failure type in each layer of the tree.

Table 1 Cluster membership of failure types after performing hierarchical clustering

Failure type	Cluster membership								
	10 clusters	9 clusters	8 clusters	7 clusters	6 clusters	5 clusters	4 clusters	3 clusters	2 clusters
F1	1	1	1	1	1	1	1	1	1
F2	2	2	2	2	2	2	2	2	2
F3	4	4	4	4	3	3	3	3	2
F4	5	5	5	5	4	4	4	3	2
F5	1	1	1	1	1	1	1	1	1
F6	6	6	6	6	5	3	3	3	2
F7	7	2	2	2	2	2	2	2	2
F8	1	1	1	1	1	1	1	1	1

F9	8	7	7	7	6	5	3	3	2
F10	1	1	1	1	1	1	1	1	1
F11	3	3	3	3	1	1	1	1	1
F12	9	8	2	2	2	2	2	2	2
F13	2	2	2	2	2	2	2	2	2
F14	3	3	3	3	1	1	1	1	1
F15	10	9	8	2	2	2	2	2	2

Table 1 shows relationships between failure types. The strength of the relationship between failure types grows with the number of tree levels in which these failure types are in the same cluster. For example, failure type F1 has a strong relationship with failure type F5, F8 and F10 because they belong to the same cluster through all levels of the hierarchical tree. As for the failure types F2, F7, F12 and F13 it applies that the strongest relationship is between failure types F2 and F13 and the relationship with failure type F7 and F12 should be repeatedly assessed.

FUTURE RESEARCH

Analyzing the failure data and finding possible relationships and similarities can be used to predict failures in the production process and control systems. This could contribute to optimizing maintenance of machines and components in the process. In future work we would like to analyze the data and find sequential patterns and to predict the failure occurrence. To be able to predict failures emerging in the production control system it is necessary not only to find relationships between failure types, but also to determine the order in which the failures occurs.

CONCLUSION

In this paper, the issue of data preparation and finding relationships using the hierarchical clustering algorithm were addressed. This data was acquired from the automotive industry and consists of records about emerging failures in the system. Each record has multiple parameters including start time and date, end time and date, duration, description, etc. To find possible relationships in this data set, hierarchical cluster analysis was chosen. To be able to perform this analysis, the data had to be preprocessed first in a way that it could serve as an input to the clustering algorithm. This preprocessing stage included data repairing, cleaning data from noise and data normalization. After the preprocessing stage, it was necessary to transform the data to a form which can be used as an input to a clustering algorithm. The proper form of input is a distance matrix. This matrix was made by calculating the Euclidean distances between the failure types, where failure type is considered as a point in the Euclidean space and its coordinates are represented by calculated parameters of the failure types. Finally, after data transformation into the distance matrix, cluster analysis with hierarchical clustering was performed. For this purpose, a software solution Statistica 13 was used. The results, in the form of hierarchical tree of clusters showed clusters of similar failure types.

Acknowledgement

This publication is the result of implementation of the project VEGA 1/0673/15: “Knowledge discovery for hierarchical control of technological and production processes” supported by the VEGA.

This publication is the result of implementation of the project: "UNIVERSITY SCIENTIFIC PARK: CAMPUS MTF STU - CAMBO" (ITMS: 26220220179) supported by the Research & Development Operational Program funded by the EFRR.

References:

1. FRIEDMAN, J. H., 2016. *Data mining and statistics: What's the connection?* Stanford: Stanford University, CA 94305. [10.11.2016] available at: http://statweb.stanford.edu/~jhf/ftp/dm-stat.pdf
2. BABCOCK, B., DATAR, M., MOTWANI, R., O'CALLAGHAN, L., 2003. Maintaining variance and k-medians over data stream windows. In: *Proc. ACM Symp. on Principles of Database Systems.*
3. DATTORRO, J., 2008. Equality relating Euclidean distance cone to positive semidefinite cone. Linear Algebra and its Applications, 428, 2597–2600. [10.11.2016] available at: https://ccrma.stanford.edu/~dattorro/EDM.pdf
4. NAZARI, Z., et all., 2015. A New Hierarchical Clustering Algorithm. In: *ICIIBMS 2015, Track2: Artificial Intelligence, Robotics, and Human-Computer Interaction*. Okinawa, Japan [10.11.2016].
5. ALPYDIN, E., 2010. *Introduction to Machine Learning*. The MIT Press, pp. 143-158.
6. FAYYAD, U. et al., 1996. *From Data Mining to Knowledge Discovery in Databases*. American Association for Artificial Intelligence, 0738-4602.
7. KOVALERCHUK, B., VITYAEV, E., 2000. Data Mining in Finance: Advances in Relational and Hybrid Methods. Springer Science & Business Media, 2000 edition, ISBN-10: 0792378040, ISBN-13: 978-0792378044