Safety Science 106 (2018) 191–202

Contents lists available at ScienceDirect

Safety Science

journal homepage: www.elsevier.com/locate/safety

A combined approach for the analysis of large occupational accident databases to support accident-prevention decision making

Lorenzo Comberti*, Micaela Demichela, Gabriele Baldissone

SAfeR – Centro Studi su Sicurezza, Affidabilità e Rischi, Dipartimento Scienza Applicata e Tecnologia, Politecnico di Torino, Corso Duca degli Abruzzi, 24, 10129 Torino, Italy

ARTICLE INFO

Keywords:
Clustering
SOM
Accident database analysis
Accident prevention

ABSTRACT

Occupational accidents are commonly collected in large databases by National Workers Compensation Authorities and companies' safety and prevention teams. The analysis of the data can be difficult because the database elements are characterized by many parameters, which are not of a numerical nature. Data mining techniques could represent an efficient tool for the identification of useful information in large databases. In 2011, a two-level clustering method, made of SOM and numerical clustering, obtained positive results in identifying critical accident dynamics. The present research proceeds from that initial methodology.

A sensitivity analysis of the coupled clustering method was carried out.

Some improvements have been designed, and an enhanced methodology has been applied to the original case study data set, for validation purposes.

This method represents an efficient tool for the analyst that has to deal with the occupational accidents data, thanks to its capability of grouping and visualizing data in a readable and exportable outcome.

The information acquired by this method can help analysts to better address the measures to be adopted in a work environment, in order to prevent occupational accidents.

1. Introduction

Occupational accident investigation is a key area for safety management systems. The process of reporting and collecting accident occurrences in an organised database, helps companies in planning procedures for risk investigation, aimed to correct existing situations and prevent further similar incidents.

Accident dynamic data are particularly relevant for risk assessment and risk-based decision making as discussed in: Leva et al. (2012) and Demichela et al. (2014a, 2014b) for high voltage equipment; Darabnia and Demichela (2013a, 2013b) for the analysis of human and organisational factors on maintenance optimisation; Gerbec et al. (2017a, 2017b) for the design of critical operation or more in general for a total safety management in Leva et al. (2014, 2015).

In a reporting system, each accident is usually described by several parameters that provide information about the dynamic, time, place, working situation and workers involved; in the European Union accidents are classified according to the ESAW system that requires more than 20 parameters to describe a single accident. The result is a multiparameter data base of large dimension: for instance, INAIL (Italian institution for insurance against accidents at work) database annually collects more than 600,000 non-fatal accidents in different occupational domains. Management of such a large amount of information can be a demanding process, with the possibility of generating errors and bias outputs, due to the flows in traditional statistical methods. Indeed, the traditional statistical methods, like multivariate analysis or linear regression, require for their application plenty of a priori assumptions over the distribution of the variables involved; therefore they become heavy to apply and could produce non-relevant results.

Furthermore, the statistical analysis returns information that is only partially useful for enhancing the prevention procedures in the work environment, as discussed in Palamara et al. (2011).

To face the problems presented by large dimension DB, a consolidated alternative, is given by the data mining approach (Edelstein, 1999; Larose, 2005), a methodology that in the last 20 years had a wide range of applications in classifying and analysing problems: from textual analysis (Zaiane, 2003) to databases of occupational accidents (Demichela and Palamara, 2007).

In data mining several algorithms have been used. Within them, the SOM (Self Organising Map), an unsupervised learning algorithm,

Abbreviations: BMU, Best Matching Unit; DB, Database; EU, European Union; ESAW, European Statistics on Accidents at Work; INAIL, Istituto Nazionale Assicurazione contro gli Infortuni sul Lavoro; IM, Input Matrix; SOM, Self Organizing Map; SKM, SOM K-mean Method

* Corresponding author.
E-mail address: lorenzo.comberti@polito.it (L. Comberti).

https://doi.org/10.1016/j.ssci.2018.03.014
Received 2 December 2016; Received in revised form 17 October 2017; Accepted 9 March 2018

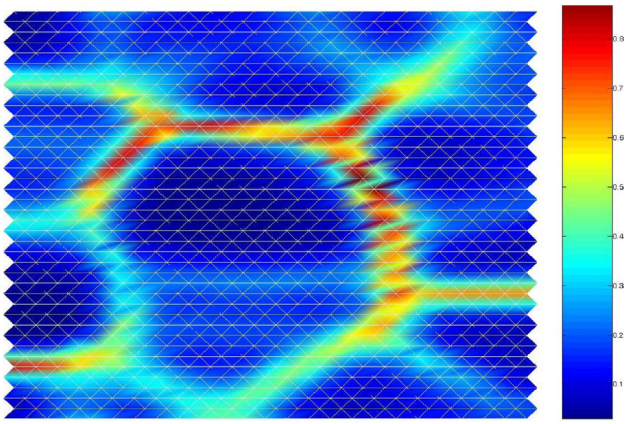

Fig. 1. SOM of an occupational accident data set.

represents a promising tool for generating a map that is preserving the original topology, from a high-dimensional data vector space to a low-dimensional map space.

SOM has already been used in many fields: web-based search applied on document classification (Kohonen et al., 2000) and web page clustering (Smith and Ng, 2003), bio informatics (Abe et al., 2006), and finance focused on quantitative analysis of debt and leasing (Sèverin, 2010) along with financial macroeconomic imbalances confrontation (López Iturriaga and Pastor Sanz, 2013).

SOM algorithms have also been applied in many risk classification problems: Liang et al. (2012) proposed SOM to classify pipeline sections with the same risk level into different risk patterns, in order to set an effective risk control strategy to prevent pipe-line damages; Asgary et al. (2012) used SOM to classify and assess the risk levels of structural fire incidents.

Since SOM provides a visualisation of input data space as it is shown in Fig. 1, it has the advantage of showing a global description of the full domain, even if it remains a qualitative description.

In order to use the analysis of previous occupational accidents as a decision-making tool for prevention purposes, it is necessary to couple the quantification procedures to the qualitative figures obtained with SOM (Demichela et al., 2014a, 2014b).

Palamara et al. (2011) proposed an innovative method, aimed at quantifying the risk associated to the clusters of occupational accidents that occurred in the Italian wood industry. The method was developed, as proposed by Vesanto and Alhoniemi (2000), in two levels of data elaboration, which combined the SOM analysis with the K-Means algorithm (McQueen, 1967), an iterative procedure that partitions the data under analysis into a certain number of clusters K.

The first level of data elaboration is applied to the coded data obtained from occupational accident DB. Through SOM, it is possible to produce a visual map of the data, following a projection process that allows the transition from the higher dimension domain of the input data space to the lower dimension domain of the projection surface. In order to better explain the process, the following example is provided: the starting point is a data set of occupational accidents where each accident is described by 8 parameters. The corresponding accident domain should be represented as a hyperspace with 8 dimensions, which has the risk of being insufficient for the analysis. By using SOM it is possible to compose the same occupational data set in a two dimensional space that is easily manageable.

SOM projection process is based on the similarity of the data, where accidents with similar characteristics are projected in next map units, while highly different accidents are projected in distant units. The map obtained is a non-metric map made of elementary units, characterised by different colours. As in Fig. 1, the colour scale of the map reflects the perturbation on the units given by the density of projected data, the blue (darker) areas represent units with a large number of data projected, while red and yellow (lighter) areas represent empty units.

The number of the blue areas can already reveal the number of groups of similar accidents, but in order to obtain an automatic clustering of the data it is necessary to apply the second stage of the method. The K-Means algorithm is applied to the numerical output deriving from the first phase, in order to produce a quantitative partition of the domain, on the basis of data similarity.

According to Palamara et al. (2011), the use of the method allows the analysts to obtain a partition of the data in clusters without any a priori assumptions about the nature of the data distribution. However, in order to positively test the effectiveness of the method, there is a need for verification of the repeatability of the results as well as the homogeneity of the clusters. The research work checked the method results with a sensitivity analysis, to introduce a revised version that is named: “SKM” (SOM K-Means Method).

Section 2 shows a more in-deep description of the method structure, its limits and the proposed improvements; Section 3 describes the results of the SKM methodology applied to the original case- study data set for validation purpose; Section 4 contains the discussion of the results, leading to the Conclusions in Section 5.

2. Methodology

The method was originally implemented in Matlab® 7.0 with an interface designed in Excel®.

In Fig. 2 the flow-chart shows the SKM logical structure.

Both SOM and K-Means algorithm manage only numbers, therefore the non-numerical information that describes each accident has to be coded into numerical information.

It is required that the pre-processing phase transforms the original data into a shape being useful for clustering: for this reason, the Accident Matrix (AM) is adopted. The matrix has a dimension “D” given by

$$D = n \times m \tag{1}$$

where “n” is the accident number and “m” is the number of variables that describes each accident.

Coding is the core of the pre-processing phase: each accident is coded from a sequence of non-numerical information to a sequence of numerical one.

As reported in Palamara et al. (2011), each parameter is coded into a numerical vector containing a sequence of 0 and a single 1; this type of coding does not affect the subsequent steps of the analysis, because it ensures that the integrity of the data distribution is not influenced by the coding process.

The complete code of each accident is given by the union of the vectors which describe the variables used for the analysis; consequently, the resulting vector will have as many “1” as the variables and as many “0” as the total number of categories less the number of variables.

The Input Matrix (IM) contains all the accidents coded in numerical vectors; its dimension is given by:

$$D_{input} = n \times p \tag{2}$$

where “n” is the number of accidents and “p” is given by the number of variables multiplied by the number of the categories used to describe them.

Let’s assume that an accident is described by 4 variables and each variable can have 5 possible different labels, then the parameter “p” will have a value of 20.

With reference to Fig. 2, after the pre-processing phase, the first level of analysis foresees the use of SOM algorithm, that firstly generates a map layer characterised by a number of units set up by the user. After this process the map layer is applied in the Input Matrix (IM), using the Hamming distance (Lourenco et al., 2002). The first level output consists of a bi-dimensional map and a numerical output that is