Application of logical analysis of data to machinery-related accident prevention based on scarce data

Sabrina Jocelyn[a,b,*], Yuvin Chinniah[b], Mohamed-Salah Ouali[b], Soumaya Yacout[b]

[a] Institut de recherche Robert-Sauvé en santé et en sécurité du travail (IRSST), 505 de Maisonneuve Blvd. West, Montreal, Quebec, Canada H3A 3C2
[b] Department of Mathematical and Industrial Engineering, Polytechnique Montréal, 2500 chemin de Polytechnique, Montreal, Quebec, Canada H3T 1J4

ARTICLE INFO

ABSTRACT

This paper deals with the application of Logical Analysis of Data (LAD) to machinery-related occupational accidents, using belt-conveyor-related accidents as an example. LAD is a pattern recognition and classification approach. It exploits the advancement in information technology and computational power in order to characterize the phenomenon under study. The application of LAD to machinery-related accident prevention is innovative. Ideally, accidents do not occur regularly, and as a result, companies have little data about them. The first objective of this paper is to demonstrate the feasibility of using LAD as an algorithm to characterize a small sample of machinery-related accidents with an adequate average classification accuracy. The second is to show that LAD can be used for prevention of machinery-related accidents. The results indicate that LAD is able to characterize different types of accidents with an average classification accuracy of 72–74%, which is satisfactory when compared with other studies dealing with large amounts of data where such a level of accuracy is considered adequate. The paper shows that the quantitative information provided by LAD about the patterns generated can be used as a logical way to prioritize risk factors. This prioritization helps safety practitioners make decisions regarding safety measures for machines.

1. Introduction

Data mining is the process of extracting hidden knowledge from data. The knowledge is extracted by means of a specific algorithm, such as support vector machine, neural networks, decision trees, association rules, or logical analysis of data (LAD). For LAD, the knowledge extracted is a set of rules or patterns describing classes of observations. In this paper, observations are accidents. The classes of observations are types of accidents, such as: a maintenance-related accident or a production-related accident. Every observation is a vector of indicators values that are recorded at the time when the accident takes place. The indicators are variables whose values describe the accident. For instance, "Presence of safeguarding" can be an indicator. It may have the value "yes" or "no" at the time of the accident. As another example, the indicator "Worker's time in current position" may have the value: 0–4 years or 5–10 years, and so on, at the time of the accident.

Data mining techniques can be preferred over traditional methods involving tests of statistical hypothesis where a minimum number of observations is an important requirement. Data mining techniques have been used for risk management in a variety of fields, including finance [1], medicine [2], transportation [3], and occupational health

and safety (OHS) [4–7]. The OHS studies deal with accidents or incidents related to workplace hazards or risk factors in general. All kinds of hazards (e.g., violence, emissions, machine-related hazards) are treated simultaneously in these studies. However, so far no study has focused on machine safety.

In OHS, Verma et al. [4] used association rules to extract frequent patterns from a database of 843 events that occurred in a plant between March 2010 and July 2013. The events were accidents, near-misses, and incidents involving material and environmental damage. Expressed as association rules, the patterns represented the acquired knowledge extracted from the database. For instance, some rules showed that behavior-related problems, such as unsafe acts performed by others, resulted in a number of injuries. Moreover, non-compliance with standard operating procedures was involved in property damage cases. The rules arrived at served as risk management tools: the variables that made them up represented the risk factors requiring risk reduction measures. For example, the factors pinpointed by the rules were useful in accident investigation. Discussions were then undertaken with safety experts that led to the identification of the root causes underlying these behavior-related problems: work stress, production pressure, overconfidence, lack of concentration, lack of training

---

for new workers, and lack of supervision for new workers. As a result, it was concluded that some measures, such as training, needed to be provided, mainly to new employees, but also to temporary workers who lacked experience.

Cheng et al. [5] made a decision tree from a database of 1542 construction accidents covering the period from 2000 to 2009. The researchers used the rules revealed by the decision tree to explain the cause-and-effect relationships. For example, one of the rules generated indicated that accidents related to the collapse of objects were more common under three concurrent conditions: (1) the source of injury was the structure and the construction facilities (e.g., scaffolding), (2) the work was performed under unsafe conditions: use of hazardous methods or procedures, and (3) the worker failed to use safeguards or ignored hazard warning signs. The rules guided preventive actions.

Silva & Jacinto [6] studied the cause-and-effect relationship regarding occupational accidents in the extraction industry. A total of 6089 accidents from the period 2005–2007 were analyzed. Three patterns were identified. Each of them characterized a specific type of accident: (1) being struck by an object, (2) physical or mental stress, (3) horizontal or vertical impact, fall of person. In order to find the patterns, a method based on multivariate analysis was applied, measuring the variables' statistical cohesion with the Pearsons' chi-square test (χ^2). The associated variables formed the patterns. The latter were used as the basis of strategies to improve safety. For example, the variables forming the pattern concerning the second type of accident focused on human behavior. Accordingly, Silva & Jacinto [6] suggested that prevention measures should be behavior-based, such as specific training sessions and well-targeted information campaigns.

Rivas et al. [7] applied various data mining techniques to determine the capacity to predict an accident or incident, and to explain such an event. The techniques tested were association rules, decision trees, Bayesian networks, support vector machine, and logistic regression. Information about the occurrence of each accident and incident was gathered by means of a survey in two companies from the mining and construction sectors respectively. The data related to the variables describing the events came from the information declared in 62 completed questionnaires, i.e., 18 accidents and 44 incidents. Rivas et al. [7] found that the best-performing predictive models were the first four above-mentioned techniques. However, only the first three demonstrated good explanatory power, showing that the occurrence of accidents in the companies could be explained by (1) task duration in hours, and (2) company contractual status (i.e., subcontractor or main contractor).

In previous OHS studies, data mining has been used for decision support to help prevent accidents. Unfortunately, the algorithms that were used inferred knowledge without covering all the observations. For instance, in association rules, the knowledge inferred is based on the identification of frequent sets of variables values in the data that meet a certain threshold. When the threshold is not satisfied, the observations concerned are rejected even though they bring new information to the database. Moreover, except for Rivas et al. [7], these studies dealt with huge databases comprising hundreds or thousands of observations. Usually, data mining techniques require large amounts of data [6] in order to extract rules describing the trends in the data. But what about plants or industries where few accidents occur? When the amount of data is limited (i.e., small sample size), the frequent sets of variables values become rarer. Accordingly, the chances of finding strong rules characterizing the data decrease. As a result, there is a need to be able to extract hidden knowledge from scarce data with adequate classification accuracy. This paper proposes to apply LAD as a data mining algorithm that is able to infer such knowledge. Indeed, LAD allows pattern generation using scarce data, as long as there is at least one observation from one class that is different from one observation from another class. Of course, when the data are too few, the patterns cannot be generalized as it was possible in [6] with thousands of data. However, the patterns can describe or

predict events only for the plants or industrial sectors concerned by the data, which is not always possible with other data mining techniques. Moreover, unlike [6] where statistical hypothesis was required, this paper proposes a study free of such hypothesis which eases the process of pattern generation. Another advantage of LAD over other data mining algorithms is the fact that all the observations are covered by patterns as long as the observations bring no contradiction into the database. Contradiction means having one class of events with an observation described with the same values of variables as an observation from another class.

Contrary to the case of Rivas et al. [7], where association rules and decision trees demonstrated high predictive performance in spite of insufficient data, these techniques performed poorly when the authors of the current paper applied them to scarce data (23×23 database) describing machinery-related accidents. For instance, only weak rules were obtained with the association rule and decision tree algorithms from Tanagra software [8]. That situation could be explained by the fact that Rivas et al. [7] might have had sufficient frequent sets of variables values in their database, which was not the case in the study reported on here. LAD has been used successfully in such diverse fields as medicine [2,9–11], finance [1], and condition-based maintenance [12,13]. LAD showed better prediction rates than other data mining techniques such as decision trees and neural networks [10,12] and was in fact the most accurate data mining technique in those cases. However, those studies dealt with huge databases, and the accuracy of LAD with small databases needed investigation.

The aim of this paper is thus two-fold. The first objective is to show that LAD is an algorithm able to characterize a narrow sample of machinery-related occupational accidents with adequate average classification accuracy. The second is to show that LAD can be used for prevention of machinery-related accidents. The use of LAD in machinery safety was suggested in a previous work [14] related to this study. The literature review from reference [14] highlighted some studies dealing with experience feedback using various data mining techniques to extract knowledge from events. One of this techniques, LAD, showed to outperform in medicine for disease diagnosis and prognosis when comes the time to distinguish and characterizing classes of events. The ability of LAD to perform on rare data was one of the main reasons why it was suggested in reference [14] for knowledge extraction suitable for machinery safety. In this paper, LAD is applied to machinery related accidents.

In the remainder of this paper, the LAD algorithm is described (Section 2) and then the application of LAD to scarce data related to machinery accidents is presented (Sections 3 and 4). The patterns generated from that application, as well as the average classification accuracy, are presented in Section 5. The potential of LAD to prevent machinery-related accidents is discussed in Section 6, based on the results.

2. Description of LAD

LAD is a data mining, pattern recognition, and classification algorithm. It is an optimization combinatorial process based on Boolean logic. LAD can deal with two-class or multi-class classification problems. In this paper, LAD is applied to a two-class problem. Generally, one of the two classes is called "positive" or "true," whereas the other is called "negative" or "false." A class can also have a precise name referring to a specific phenomenon or event, just as the example of Section 2.1. It presents two classes: (1) Accidents resulting in harm versus. (2) Accidents not resulting in harm. Moreover, as presented in Section 3, the study regarding this paper deals with the two following specific classes: (1) a maintenance-related-accident class and (2) a production-related-accident class. As with any data mining technique, LAD is performed in two main phases: (1) training, then (2) testing.

The training phase consists of browsing any numerical database in order to extract knowledge naturally hidden in the database. The