Profiling risk sensibility through association rules

Beatrice Lazzerini*, Francesco Pistolesi

Dipartimento di Ingegneria dell'Informazione: Elettronica, Informatica, Telecomunicazioni, University of Pisa, Largo Lucio Lazzarino 1, 56122 Pisa, Italy

Abstract

In the last recent years several approaches to risk assessment and risk management have been adopted to reduce the potential for specific risks in working environments. A safety culture has also developed to let workers acquire knowledge and understanding of risks and safety. Notwithstanding, risks still exist in every workplace. One effective way to improve workers' sensibility to risk, i.e., their ability to effectively assess and control the risks they are exposed to, is risk management training. Unfortunately, people may perceive risks in different ways depending on subjective assessment of the characteristics and severity of the considered risks, and may have tendencies to either take or avoid actions that they feel are risky. Therefore, the knowledge of how workers assess each of the risks they may be exposed to in the workplace is a key factor to conceive effective custom risk management training. In this paper we present a novel approach, based on association rules, to workers' profiling with respect to risk perception and risk propensity in order to provide each of them with specific customized risk management training.

Keywords: Association rules, Frequent patterns, Risk perception, Risk propensity

1. Introduction

Many modern working environments are characterized by a huge quantity of risks the workers are exposed to. When workers do their job, or they simply stay in the workplace, negative events may happen, whose effects can damage their health and safety. Actually, in the last decades safety management has become an important field of study, much emphasis has been placed on the need for a deep understanding of risk management concepts and principles (Aven, 2012), more and more powerful tools have been developed for risk analysis and management (Dubois, 2010) (Lazzerini & Mkrtchyan, 2010) (Lazzerini & Mkrtchyan, 2011) (Misra, 2008). At the same time, industrial systems and machineries have been redesigned and provided with efficient safety features, and working environments have evolved into safer places (ISO31000, 2009) (Leitch, 2010): despite this, residual risks remain unacceptably high (ISO27001, 2005).

Workers exposed to risks should be aware of the nature of these risks so as to deal with them as well as appropriate. Human subjectivity causes people to behave differently in similar risky situations: the two main variables playing an important role in the interaction between

*Corresponding author. Tel.: +39 050 2217558; fax: +39 050 2217600.

Email addresses: `b.lazzerini@iet.unipi.it` (Beatrice Lazzerini), `f.pistolesi@iet.unipi.it` (Francesco Pistolesi)

a person and a possibly risky environment are risk perception and risk propensity (Keil et al., 2000). *Risk perception* is the subjective way with which a person estimates characteristics and gravity of hazardous situations (Bouyer et al., 2001) (Chauvin et al., 2007) (Peters & Slovic, 1996) (Sjöberg, 2000) (Slovic, 1987), while *risk propensity* is a person's tendency to take or avoid risks (Sharma et al., 2009) (Sitkin & Weingart, 1995) (Sueiro et al., 2011).

Risk perception is influenced by a variety of factors, including past experience and knowledge, past health status, psychological, social, political, and cultural factors, mood and emotions, personal knowledge about the risky condition, trust in risk management institutions, age, sex, locus of control (Horswill & McKenna, 1999) (Klein & Helweg-Larsen, 2002), optimism bias (Costa-Font et al., 2009) (Klein & Helweg-Larsen, 2002), etc. Similarly, risk propensity is influenced by diverse factors as personality and experience, cultural background, mood, feelings, gender, education, job position, age, etc.

Although extensive research has been made about the factors that determine, respectively, risk perception and risk propensity, no general model exists to explain the diverse behavioural strategies in dealing with a given risk. Further, little is known about the interrelations that exist among risk perception, risk propensity, and decisions involving risk (Keil et al., 2000).

For this reason, training policies have been purposely devised and adopted with the aim of improving peoples' ability of quickly identifying a source of danger and its potential dangerous effects. The main objective of such a training process is increasing the risk sensibility and awareness level of the workers in order to obtain a safer interaction with the risk itself (Nicholson et al., 2001). Unfortunately, when training is addressed to a heterogeneous group of workers, some workers may obtain a sufficient level of risk awareness, while others may maintain an inappropriate way of interaction with the risk. Therefore, the training process should be tailored to the specific risk sensibility profile of the involved worker.

In this paper we describe a way for customizing the training process for risk awareness by adapting it to the worker's risk perception and risk propensity, in order to obtain the best result in terms of learning. In particular, we consider some criticality factors whose correlation with risk perception and risk propensity has been stated by sociology and psychology experts, namely, gender, age, level of education, income, risk knowledge, work control at work site, professional role, injury frequency, effect seriousness, delayed occurrence of effects, role repetitiveness, industrial injuries and diseases, acquired skills, perception of risk control, work gratification, state of health, safety culture in the company, anxiety level, self-esteem, worry level.

Our choice of the criticality factors is essentially based on heuristic considerations as our main objective is to prove the feasibility and effectiveness of the proposed association rule-based approach to risk sensibility profiling.

To achieve our objective, we deduce the risk sensibility profile of a worker from his/her behaviour in dealing with one or more risks, and from the set of criticality factors introduced above. More precisely, given a set of risks to which a group of workers is exposed, we consider, for each risk, a set of actions a worker should or may perform to prevent the risk to occur. Then each worker is interviewed by asking him/her what action (or actions) he/she would perform to prevent the occurrence of each risk to which he/she is exposed. Each worker is also asked to answer some questions related to the criticality factors. After collecting a sufficient number of interviews, through a data mining process we look for association rules (Agrawal et al., 1993a) (Agrawal et al., 1993b) concerning risks, prevention actions and criticality factors. Once the association rules have been generated, each of those associated with a high level of interestingness, expressed in terms of appropriate indexes (Agrawal & Srikant, 1994) (Wu et al., 2010), may represent a risk sensibility profile. In this way, a risk sensibility profile consists in a particular configuration of the criticality factors and a specific behaviour towards one or more risks. We consider both single-risk profiles and multi-risk profiles.

This work represents a first step towards significantly improving risk management as we provide the risk manager with direct knowledge of the risk sensibility profiles of the workers that will be the target of training aimed at risk awareness.

The rest of this paper is organized as follows. In Section 2 we introduce the basics of association rule mining; in Section 3 we introduce the concepts of single-risk profile and multi-risk profile, and detail our association rule-based approach to risk sensibility profiling; in Section 4 we give an illustrative example of the application of the proposed risk sensibility profiler. Finally, we provide concluding remarks in Section 5.

2. Data mining

Data mining is the process of discovering interesting correlations or useful patterns in large data sets. One of the most important data mining techniques is *association rule mining*, first introduced in (Agrawal et al., 1993b), which aims to discover all significant associations between items in data repositories.

2.1. Frequent patterns and association rules

Frequent patterns are itemsets, structures or sequences of items which are recurrent in a dataset, i.e., they appear together with frequency higher than a specified threshold (Han et al., 2007). A frequent pattern is called *frequent itemset*. Finding frequent patterns allows to discover associations and correlations among the items in a dataset.

Let $\mathcal{U} = \{i_1, i_2, \ldots, i_m\}$ be a set of items. A *transaction* T is a subset of $\mathcal{U}$. A set of items $X \subseteq \mathcal{U}$ is called *itemset*; an itemset which contains k items is also named a *k-itemset*. A transaction T contains an itemset X if $X \subseteq T$. Let $A \subset \mathcal{U}$ and $B \subset \mathcal{U}$ be two itemsets such that $A \cap B = \emptyset$. An *association rule* is an implication of the form $A \Rightarrow B$, with A and B called, respectively, *antecedent* and *consequent*. The two most used measures of the importance of an association rule are support and confidence. The *support* expresses the proportion of transactions in the transaction set containing $A \cup B$. The event of finding both sets A and B occurs with a probability $P(\mathcal{A} \cap \mathcal{B})$, where $\mathcal{A}$ and $\mathcal{B}$ are, respectively, the event of finding the itemset A and the event of finding the itemset B in a transaction. Hence the support (abbreviated as supp) of a rule $A \Rightarrow B$ is defined as

$$\mathrm{supp}(A \Rightarrow B) = P(\mathcal{A} \cap \mathcal{B}). \tag{1}$$

The support is also indicated with $\mathrm{supp}(A \cup B)$.

The *confidence* (abbreviated as conf) of the rule $A \Rightarrow B$ is defined as the percentage of transactions containing A that also contain B:

$$\mathrm{conf}(A \Rightarrow B) = P(\mathcal{B}|\mathcal{A}) = \frac{\mathrm{supp}(A \cup B)}{\mathrm{supp}(A)}, \tag{2}$$

where $P(\mathcal{B}|\mathcal{A})$ is the conditional probability.

During the rule mining process, only the rules having support and confidence greater than user-defined thresholds are considered: these rules are called *strong rules*.

The association rule generation can be described as a two-step process (Agrawal et al., 1993b):

1. *large itemset generation*
   find all itemsets having support above the chosen minimum support s_{min}; these itemsets are called *large* itemsets;
2. *rule generation*
   use the large itemsets to generate the desired rules. Each generated rule has to satisfy also a minimum confidence.

The association rules can be generated starting from the set $\mathcal{L}$ of large itemsets returned by step 1, as follows:

- for each large itemset $l \in \mathcal{L}$, generate all its non-empty subsets;
- for each non-empty subset $s \subset l$, consider the rule $s \Rightarrow (l - s)$ if and only if:

$$\frac{\text{supp}(l)}{\text{supp}(s)} \geq c_{min}$$

  where c_{min} is the minimum confidence.

Since all the rules are generated from large itemsets, they automatically satisfy the minimum support conditions, so only the minimum confidence level has to be checked.

A lot of algorithms for mining association rules use criteria based on support and confidence. Although these approaches are able to avoid the generation of a high number of uninteresting rules, many of the generated rules may still be not interesting to the user. This may happen, e.g., when a too low support threshold is used. To solve this problem further interestingness measures can be taken into account, based on statistical significance and correlation analysis. Several interestingness measures have been proposed; one of the most popular is, e.g., the *lift*, which is defined as

$$\text{lift}(A \Rightarrow B) = \frac{P(\mathcal{A} \cap \mathcal{B})}{P(\mathcal{A})P(\mathcal{B})} = \frac{\text{conf}(A \Rightarrow B)}{\text{supp}(B)}. \tag{3}$$

Unfortunately, the lift is not a good measure for correlations, especially in large transactional databases, as it is affected by the number of the so-called *null transactions*, i.e., transactions that do not contain any of the itemsets under consideration. An effective *null-invariant* measure is the *Kulczynski index* (Bradshaw, 2001) (Kulczynski, 1927) (abbreviated as Kulc) defined as the average of two conditional probabilities:

$$\text{Kulc}(A \Rightarrow B) = \frac{1}{2}\left(P(\mathcal{B}|\mathcal{A}) + P(\mathcal{A}|\mathcal{B})\right). \tag{4}$$

The Kulc index can be interpreted as the average of two confidence measures, i.e., the confidences of $A \Rightarrow B$ and $B \Rightarrow A$. More precisely,

$$\text{Kulc}(A \Rightarrow B) = \frac{\text{supp}(A \cup B)}{2}\left(\frac{1}{\text{supp}(A)} + \frac{1}{\text{supp}(B)}\right). \tag{5}$$

3. Risk sensibility profiles

A *risk sensibility profile* consists in a specific behaviour towards one or more risks and a particular configuration of the criticality factors. Generally, workers exposed to the same risk (or risks) are characterized by a personal way of thinking about these risks and by a specific behaviour towards them. In the following, we will take both single-risk profiles and multi-risk profiles into account.

3.1. Single-risk profiles

Let us consider an environment characterized by a risk r associated with a set of actions which can be performed to prevent its occurrence.

Supposing that a group of workers are exposed to the risk r, each worker will choose one or more actions to protect himself/herself from the risk and he/she will potentially do this in different ways.

In this scenario an item can be a risk, an action or a criticality factor. A transaction (i.e., an interview) is composed by a risk, a set of actions and a specific configuration of all the criticality factors. Each criticality factor can assume one of a set of predefined values.

Let $\mathcal{T} = \{t_1, t_2, \ldots, t_n\}$ be the set of transactions we consider as the dataset, let $\mathcal{A} = \{a_1, a_2, \ldots, a_m\}$ be the set of actions preventing the risk r to occur, let $C = \{c_1, c_2, \ldots, c_k\}$ be the set of criticality factors, and let $\mathcal{V}^{c_i} = \{v_1^{c_i}, v_2^{c_i}, \ldots, v_{h_{c_i}}^{c_i}\}$ be the set of values of criticality factor c_i. To avoid unnecessary use of notation, and since this does not cause ambiguity, we will make the following assumption: we will suppose that all criticality factors can assume the same values, and we will denote the set of values of all criticality factors as $\mathcal{V} = \{v_1, v_2, \ldots, v_h\}$, thus omitting some indexes. Let us observe that it is easy to generalize the previous assumption to allow different criticality factors to assume different values in a real case. For each transaction $t \in \mathcal{T}$, we define a $k \times h$ matrix Q^t:

$$Q^t = \begin{pmatrix} q_{11} & \cdots & q_{1h} \\ \vdots & \ddots & \vdots \\ q_{k1} & \cdots & q_{kh} \end{pmatrix}$$

$$\text{s.t. } q_{ij} = Q^t_{ij} = \begin{cases} 1 & \text{if } c_i \text{ has the value } v_j \\ 0 & \text{otherwise.} \end{cases}$$

Note that a criticality factor can assume only one value for each transaction, hence $\sum_{\lambda=1}^{h} q_{i\lambda} = 1$, where $i = 1, 2, \ldots, k$.

We will refer to the set of actions with which a worker prevents the risk r as a *strategy*. A strategy is represented by an m-dimensional vector s where the i-th component s_i is equal to one or zero depending on the fact that the worker would perform or not the i-th action to prevent r. For instance, a strategy based on the actions a_1, a_3 and a_{m-1} is represented by the vector $(1, 0, 1, 0, \ldots, 0, 1, 0)$.

Given a transaction t we define the set $\mathcal{J}_s^t$ containing the indexes of the actions a worker would perform to prevent the risk r as:

$$\mathcal{J}_s^t = \bigcup_{s_i=1} i \quad \text{with } i = 1, 2, \ldots, m. \tag{6}$$

Finally, we indicate with $\mathcal{J}_C^t$ the set of pairs, each composed by the index of a criticality factor and the index of the associated value in the transaction t:

$$\mathcal{J}_C^t = \bigcup_{j=1}^{k} \bigcup_{\lambda=1}^{h} (j, \lambda) \quad \text{s.t. } Q^t_{j\lambda} = 1. \tag{7}$$

A transaction $t \in \mathcal{T}$ can now be defined as:

$$t = \{r, \bigcup_{i \in \mathcal{J}_s^t} a_i, \bigcup_{(j,\lambda) \in \mathcal{J}_C^t} (c_j, v_\lambda)\}. \tag{8}$$

So the itemsets are all the sets having the following structure:

$$I = \{r, \bigcup_{i \in \mathcal{J}'_s} a_i, \bigcup_{(j,\lambda) \in \mathcal{J}'_C} (c_j, v_\lambda)\} \tag{9}$$

such that $\mathcal{J}'_s \subseteq \mathcal{J}^t_s$, $\mathcal{J}'_C \subseteq \mathcal{J}^t_C$.

Let $\mathcal{I}$ be the set of all the itemsets that can be generated from a given transaction. The support of an itemset $I \in \mathcal{I}$ is expressed by:

$$\text{supp}(I) = \frac{\sum_{j=1}^{n} \alpha_j}{n} \quad \text{s.t.} \quad \alpha_j = \begin{cases} 1 & \text{if } t_j \supseteq I \\ 0 & \text{otherwise} \end{cases}, \tag{10}$$

where $j = 1, 2, \ldots, n$ and $t_j \in \mathcal{T}$ is the j-th transaction. The itemsets having a good support have to be considered as input for the rule generation step. If s_{min} is the minimum support, the set of large itemsets $\mathcal{L}$ is

$$\mathcal{L} = \{I \in \mathcal{I} \ \text{ s.t. } \ \text{supp}(I) \geq s_{min}\}. \tag{11}$$

We are interested in association rules having a risk and a subset of criticality factors as antecedent and a subset of actions as consequent. In this way we can study the logical connection strength between a risk and some criticality factors on one hand, and a set of actions on the other hand.

Given a large itemset $l \in \mathcal{L}$ having the structure shown in (9), we will consider association rules of the form:

$$R : \{r, \bigcup_{(j,\lambda) \in \mathcal{J}'_C} (c_j, v_\lambda)\} \Rightarrow \bigcup_{i \in \mathcal{J}'_s} a_i. \tag{12}$$

A rule like (12), called *single-risk rule*, means that "*if a worker is exposed to the risk r and he/she is characterized by the criticality factor values in* $\bigcup_{(j,\lambda) \in \mathcal{J}'_C} (c_j, v_\lambda)$*, then he/she will perform the actions in* $\bigcup_{i \in \mathcal{J}'_s} a_i$ *to prevent the risk r to occur*".

When all the rules from the large itemsets in $\mathcal{L}$ have been generated, we have to analyse their confidence and Kulc index. In fact the only support evaluation is not enough to assume that a rule represents a profile. Let c_{min} and k_{min} be the minimum confidence and the minimum Kulc index, respectively. Using the equations (2) and (5), we are able to maintain only *strong* rules such that:

$$\begin{cases} \dfrac{\text{supp}(\{r, \bigcup_{i \in \mathcal{J}'_s} a_i, \bigcup_{(j,\lambda) \in \mathcal{J}'_C} (c_j, v_\lambda)\})}{\text{supp}(\{r, \bigcup_{(j,\lambda) \in \mathcal{J}'_C} (c_j, v_\lambda)\})} \geq c_{min} \\[2ex] \dfrac{\text{supp}(\{r, \bigcup_{i \in \mathcal{J}'_s} a_i, \bigcup_{(j,\lambda) \in \mathcal{J}'_C} (c_j, v_\lambda)\})}{2} \Bigg(\dfrac{1}{\text{supp}(r, \bigcup_{i \in \mathcal{J}'_s} a_i)} + \\[2ex] \qquad \dfrac{1}{\text{supp}(\{r, \bigcup_{(j,\lambda) \in \mathcal{J}'_C} (c_j, v_\lambda)\})} \Bigg) \geq k_{min}. \end{cases} \tag{13}$$

Now we can give the following definition for a *single-risk profile*:

Definition 3.1. Let r be a risk, let $\mathcal{W} \subseteq C$ be a set of criticality factors and let $\mathcal{A}^\star \subseteq \mathcal{A}$ be a subset of the actions a worker would perform to prevent r to occur. A *single-risk profile* $\mathcal{P}_r$ is a strong association rule having the structure $\{r, \mathcal{W}\} \Rightarrow \mathcal{A}^\star$.

A single-risk profile $\mathcal{P}_r$ is represented by the specific correlation between a risk r and the values of a subset of the criticality factors on one hand, and a subset of the actions avoiding the risk r to occur on the other hand.

Setting the minimum level of interestingness measures is a critical operation because there are situations where high thresholds could cause information loss. On the other hand, low thresholds can generate a too high number of uninteresting rules. In our approach, the support is the most critical indicator.

Usually, the support threshold is set to values of at least 70% when frequent patterns have to be privileged. For example, in a *Market Basket Analysis* context, the support has to be high because only an association that occurs a lot of times might generate a potentially strong association rule.

On the other hand, in risk sensibility profile mining we might need to take into account also lower support thresholds in order to identify niche profiles.

A *niche profile* is a risk sensibility profile describing behavioural choices which, although not so common, are anyhow important to collect.

3.2. Multi-risk profiles

The single-risk mining approach is the simplest way to find out risk sensibility profiles connected to a risk. Unfortunately, in a working context there may be more risks to which a worker can be contemporaneously exposed, depending on the working environment nature. In fact, when workers do their task (even a simple task) they generally use a machine tool, some utensils and they often have to move towards different places. In order to take these situations into account,

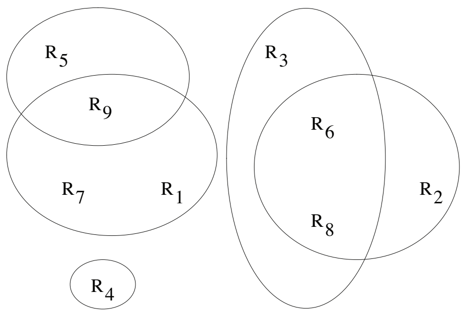

Fig. 1: Risk contexts (ovals) in a working environment including nine risks

the previously presented approach has to be enriched to include the possibility that more risks could occur simultaneously. From now on, we will refer to a working environment in which one or more risks may occur as a *risk context*. A risk context is linked to a specific task. In Fig. 1 a working environment with some risk contexts is shown. In the figure, we can notice that the same risk may be included in more than one risk context.

We can distinguish two kinds of risk context depending on the nature of the involved risks:

- *homogeneous* risk context, in which the risks have the same nature;
- *heterogeneous* risk context, characterized by risks of different nature.

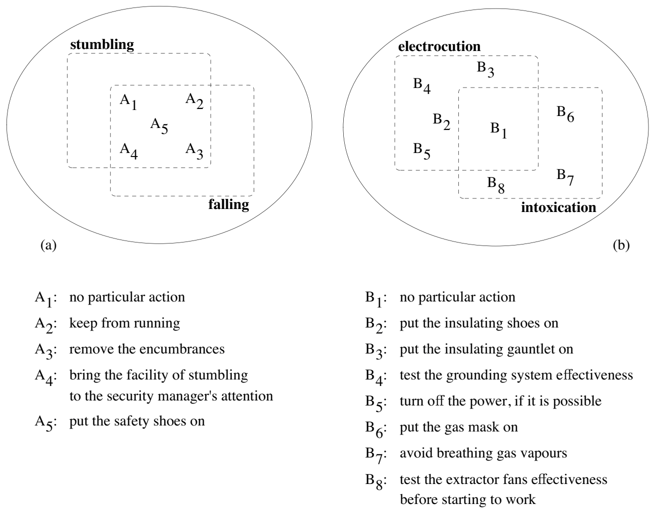

Fig. 2: Two risk contexts (the dotted rectangles include the prevention actions associated with the considered risks): (a) homogeneous risk context; (b) heterogeneous risk context

Risks of the same nature are characterized by the same set of actions preventing them to occur: these risks are caused by the same factors and they expose a worker to the same damage. For instance, two risks of this kind are the stumbling risk and the falling risk, which could be prevented using the same actions. On the other hand, typically, risks of different nature do not share (all) actions, that is, each risk has its own set of preventing actions. Think, e.g., of a context in which the electrocution risk and the intoxication risk could occur together. In such a context, a worker should use different kinds of actions to prevent both risks. In Fig. 2 an example of homogeneous and heterogeneous risk contexts is shown.

In order to perform an association rule mining process in a multi-risk scenario, we have to make some changes to the model we have proposed in the previous section. In fact, when the single-risk approach is used, only one risk at a time is considered, hence during the mining process there is no ambiguity regarding the link between a risk and its preventing actions. On the contrary, in a multi-risk context there is the need of connecting a risk with its preventing actions. A multi-risk association rule has to clearly explain what risk an action refers to, since sometimes the same action could be used to prevent different risks. Let $\mathcal{R} = \{r_1, r_2, \ldots, r_n\}$ be the set of risks which could occur in the considered context. Let $\mathcal{A}^\star = \{a_1, a_2, \ldots, a_m\}$ be the set of actions preventing all the risks in $\mathcal{R}$. Of course, each action may be linked to one or more risks. Let $\mathcal{A}^r \subseteq \mathcal{A}^\star$ be the set of actions that can be performed to prevent a risk $r \in \mathcal{R}$ to occur.

Definition 3.2. A *risk-action item* is a pair (r, A_i^r) meaning that the risk $r \in \mathcal{R}$ should be prevented using the action $A_i^r \in \mathcal{A}^r$.

Let $\mathcal{T}$ be the set of transactions we consider as the dataset, and let $C = \{c_1, c_2, \ldots, c_k\}$ be the set of criticality factors. Finally, let $\mathcal{V} = \{v_1, v_2, \ldots, v_h\}$ be the set of values the criticality factors can assume (in this case too, for the sake of simplicity, we suppose that all criticality factors can assume the same values).

Let us define an $n \times m$ matrix S, called *strategy matrix*, to represent what are the actions preventing each risk $r_i \in \mathcal{R}$:

$$S = \begin{pmatrix} s_{11} & \cdots & s_{1m} \\ \vdots & \ddots & \vdots \\ s_{n1} & \cdots & s_{nm} \end{pmatrix} \text{ s.t. } s_{ij} = \begin{cases} 1 & \text{if } r_i \text{ can be prevented} \\ & \text{by performing } a_j \\ 0 & \text{otherwise.} \end{cases}$$

In order to perform multi-risk association rule mining, first of all we have to pre-process the set $\mathcal{A}^\star$ to obtain a set $\mathcal{A}$ whose elements are risk-action items. Using the previous matrix S we can define $\mathcal{A}$ as:

$$\mathcal{A} = \bigcup_{(i,j)|s_{ij}=1} (r_i, A_j^{r_i}) \tag{14}$$

where $i = 1, 2, \ldots, n$ and $j = 1, 2, \ldots, m$.

Consider the following sets of indexes

$$\begin{aligned} \mathcal{J}_\mathcal{R} &= \{1, 2, \ldots, n\}, \\ \mathcal{J}_\mathcal{A} &= \textstyle\bigcup_{s_{i,j}=1}(i, j), \\ \mathcal{J}_C &= \{1, 2, \ldots, k\}, \end{aligned}$$

which contain the indexes of the risks, of the risk-action items, and of the criticality factors, respectively. In a multi-risk context a transaction $t \in \mathcal{T}$ has the following form:

$$t = \{\bigcup_{i \in \mathcal{J}_\mathcal{R}} r_i, \bigcup_{(i,j) \in \mathcal{J}'_\mathcal{A}} (r_i, A_j^{r_i}), \bigcup_{(j,\lambda) \in \mathcal{J}^t_C} (c_j, v_\lambda)\} \tag{15}$$

where $\mathcal{J}'_\mathcal{A} \subseteq \mathcal{J}_\mathcal{A}$. We recall that $\mathcal{J}^t_C$ contains pairs of indexes (j, λ) relative to the matrix Q^t defined in Section 3.1, where j is the row index pertinent to a criticality factor, and λ is the column index of the criticality factor value relative to the worker represented by the transaction t.

An itemset in a multi-risk context has the form:

$$I = \{\bigcup_{i \in \mathcal{J}'_\mathcal{R}} r_i, \bigcup_{(i,j) \in \mathcal{J}''_\mathcal{A}} (r_i, A_j^{r_i}), \bigcup_{(j,\lambda) \in \mathcal{J}'_C} (c_j, v_\lambda)\} \tag{16}$$

where $\mathcal{J}'_\mathcal{R} \subseteq \mathcal{J}_\mathcal{R}$, $\mathcal{J}''_\mathcal{A} \subseteq \mathcal{J}'_\mathcal{A}$ and $\mathcal{J}'_C \subseteq \mathcal{J}^t_C$.

Based on the itemsets like (16) we will consider two categories of association rules to describe different implications which could exist in a multi-risk environment. More precisely, these two categories of association rule express two kinds of multi-risk profile which we will describe in the next paragraphs: first-level multi-risk profiles and second-level multi-risk profiles.

3.2.1. First-level multi-risk profiles

The first kind of rule we are interested in is an implication having the following structure:

$$\{\bigcup_{i \in \mathcal{J}'_\mathcal{R}} r_i, \bigcup_{(j,\lambda) \in \mathcal{J}'_C} (c_j, v_\lambda)\} \Rightarrow \bigcup_{(i,j) \in \mathcal{J}''_\mathcal{A}} (r_i, A_j^{r_i}) \; . \tag{17}$$

A rule like (17) means that "*if a worker is exposed to the risks $\bigcup_{i \in \mathcal{J}'_\mathcal{R}} r_i$ and is characterized by criticality factors $\bigcup_{(j,\lambda) \in \mathcal{J}'_C} (c_j, v_\lambda)$, then this worker will perform the actions $\bigcup_{(i,j) \in \mathcal{J}''_\mathcal{A}} (r_i, A_j^{r_i})$ to prevent the risks he/she is exposed to*".

We indicate this kind of rule as a *first-level multi-risk rule*. We can note that the consequent of a first-level multi-risk rule contains the conjunction of different behaviours, each related to a given risk, all triggered by the same criticality factors of a person exposed to the risks in the antecedent. If the rule (17) has good values of the interestingness parameters (i.e., good support, confidence and Kulc index) then it is a strong multi-risk rule and it can be associated with a *first-level multi-risk profile*.

3.2.2. Second-level multi-risk profiles

In general, not all the workers that present the same value of their criticality factors and are exposed to the same risks show the same behaviours towards all the risks. Actually, we can often find out that these workers can be classified into two or more groups, based on a similar behaviour towards a subset of the involved risks. In practice, the set of all the risks may be partitioned into disjoint subsets. Stated in other terms, this means that the specific behaviour a worker has towards a risk (or risks) could influence his/her way of preventing other risks. For instance, the actions a worker performs to prevent a falling could be connected to specific actions he/she performs to protect himself/herself from electrocution.

An association rule which expresses this typology of profile has the form:

$$\{\bigcup_{i\in\mathcal{J}'_{\mathcal{R}}} r_i, \bigcup_{(j,\lambda)\in\mathcal{J}'_{C}} (c_j, v_\lambda), \bigcup_{(i,j)\in\mathcal{J}^{\alpha}_{\mathcal{A}}} (r_i A^{r_i}_j)\} \Rightarrow \bigcup_{(i,j)\in\mathcal{J}^{\beta}_{\mathcal{A}}} (r_i A^{r_i}_j) \tag{18}$$

where $\mathcal{J}^{\alpha}_{\mathcal{A}}, \mathcal{J}^{\beta}_{\mathcal{A}} \subset \mathcal{J}''_{\mathcal{A}}$ and $\mathcal{J}^{\alpha}_{\mathcal{A}} \cap \mathcal{J}^{\beta}_{\mathcal{A}} = \emptyset$.

A rule like (18) means that "*if a worker is exposed to the risks $\bigcup_{i\in\mathcal{J}'_{\mathcal{R}}} r_i$, is characterized by criticality factors $\bigcup_{(j,\lambda)\in\mathcal{J}'_{C}} (c_j, v_\lambda)$, and prevents the risks in $\mathcal{J}^{\alpha}_{\mathcal{A}}$ by performing the actions $\bigcup_{(i,j)\in\mathcal{J}^{\alpha}_{\mathcal{A}}} (r_i A^{r_i}_j)$ then he/she will perform the actions $\bigcup_{(i,j)\in\mathcal{J}^{\beta}_{\mathcal{A}}} (r_i A^{r_i}_j)$ to prevent the risks in $\mathcal{J}^{\beta}_{\mathcal{A}}$*".

If a rule like (18) is characterized by good levels of interestingness (i.e., good support, confidence and Kulc index) it can be associated with a *second-level multi-risk profile*.

4. Experiments

In order to apply and test the approach proposed in this paper, we have implemented a prototype of the risk sensibility profiler. The profiler has been implemented in Java, using the JEE platform. In particular, the system is a three-tier application including i) a back-end tier consisting of a relational database managed by a MySQL DMBS; ii) a middle tier containing an EJB-based object-relational mapping, and the business logic for managing the operations on the data; iii) a front-end tier consisting in a web interface based on Java servlets and Java server pages. Thanks to this approach, a remote collecting process has been made possible, allowing the workers to anonymously take part to the experiment through the website.

We have used data collected by interviewing voluntary workers, due to restrictions the law establishes on privacy of personal behaviour. For this reason, the dataset we have populated is not so large as a data mining technique would require. However, with a number of 80 transactions we have been equally able to prove the risk sensibility profiler effectiveness.

We have considered a group of five risks, all typical of the shoe factory environment (namely, the cut, stumbling, falling, electrocution and intoxication risks); for each of these risks we have proposed a set of preventive actions (including, for the sake of completeness, no action at all). The workers we interviewed are exposed to one or more of the whole set of risks doing their job. The interviews have been stored in a database and then they have been given as input to the profiler. We have experimented both the single-risk and the multi-risk profiling process.

The single-risk profiling experiment has been carried out considering all the risks, one at a time. For the multi-risk experiment, we have grouped the risks in subsets, each representing a specific risk context. We will show the results we have achieved by testing the system with one risk for the single-risk procedure, and with one heterogeneous risk context for the multi-risk approach.

For the single-risk procedure we consider the *cut* risk. The actions we have proposed to prevent this risk are:

- activation of the machinery safety elements = ASE;
- verification of the safety elements efficiency = VSE;
- put the gauntlet on = PGO;
- keep hands away from the cutting elements = KHA;
- switch off the cutting machine to fix a fault = SOF;
- periodically check and sharpen the cutting utensils = PCS;
- no particular action = NPA.

The criticality factors we have considered have been listed in Section 1. In a transaction (i.e., an interview), each factor assumes one value in the set of linguistic labels {*very low, low, medium, high, very high*}, except four factors. Precisely, the *g*ender and *p*rofessional role factors are categorical variables (e.g., *g*ender may be either *male* or $female$), while *a*ge and *i*ndustrial injuries and diseases are integer variables.

Based on heuristic considerations, we have set a minimum support of 0.2, a minimum confidence of 0.9 and a minimum Kulc index of 0.7. Once the mining process has completed, it has returned the following strong association rules, each one expressing a single-risk profile:

1. gender = *male*, self-esteem = *high* $\Rightarrow$ PGO;
2. anxiety level = *medium*, acquired skills = *high* $\Rightarrow$ ASE, PGO, SOF;
3. age < 25, state of health = *high* $\Rightarrow$ NPA;
4. risk knowledge = *high*, perception of risk control = *high* $\Rightarrow$ KHA, PCS;
5. industrial injuries and diseases > 2 $\Rightarrow$ ASE, VSE, PGO, KHA, PCS.

The interestingness parameter values for the five rules are summarized in Table 1. The set of profiles we have discovered for the cut risk shows different ways the workers behave to prevent it, depending on (some of) their criticality factors.

Once the workers have been classified into one of the previous profiles, we have a characterization of their way of interaction with the considered risk. This information will be provided to the person in charge of the risk awareness training task who can customize the learning process according to the profile of the person he/she deals with.

We can observe that using the previous single-risk profiles we cannot achieve an exhaustive classification of the workers, due to the chosen minimum support threshold. Indeed, we

Table 1: Interestingness parameter values of the mined strong rules

rule ID	support	confidence	Kulc
1.	0.41	0.98	0.85
2.	0.31	0.97	0.94
3.	0.49	0.94	0.98
4.	0.23	1	0.74
5.	0.37	0.99	0.85

have been able to classify 80% of the workers. This percentage could however be increased by adopting a support threshold closer to zero.

Regarding multi-risk profiling, we have considered a heterogeneous risk context including the *intoxication* risk (IR) and the $falling$ risk (FR). In order to prevent the intoxication risk we have proposed the following actions:

- activate the extractor fans = AEF;
- put the gas mask on = PGM;
- avoid to breathe during gas emission = ABG;
- no particular action = NPA.

The falling risk can be prevented performing one or more of the following actions:

- use of the safety snap hooks = SSH;
- check of the platform/ladder stability = PLS;
- keep from climbing = KFC;
- keep from moving rapidly = KMR;
- no particular action = NPA.

In order to improve rule interpretation, and thanks to the fact that we have chosen a risk context whose risks do not share any actions (apart from the obvious action NPA, which will be indicated as NPA_{IR} if performed on the intoxication risk and as NPA_{FR} if performed to prevent the falling risk), we will list the multi-risk association rules without explicitly associating the actions with their own risk: this does not cause ambiguity.

Setting a support threshold of 0.2, a confidence threshold of 0.9 and a Kulc index threshold of 0.7 we have collected the following strong rules:

1. level of education = *low*, worry level = *medium* $\Rightarrow$ PGM, KMR;
2. perception of risk control = *high*, state of health = *high* $\Rightarrow$ AEF, NPA_{FR};
3. acquired skills = *high* $\Rightarrow$ AEF, PGM, SSH, PLS, KFC;
4. gender = *male*, age < 30, worry level = *low* $\Rightarrow$ ABG, KMR;
5. perception of risk control = *high*, gender = *male*, PGM $\Rightarrow$ SSH, PLS;

6. work gratification = *low*, KFC $\Rightarrow$ NPA_{IR}.

Table 2: Interestingness parameter values of the multi-risk rules

rule ID	**support**	**confidence**	**Kulc**
1.	0.52	0.99	0.94
2.	0.35	1	0.77
3.	0.27	0.95	0.72
4.	0.44	1	0.9
5.	0.38	0.96	0.83
6.	0.35	1	0.89

As we can see, the multi-risk experiment has returned a group of four first-level profiles (from 1 to 4) and two second-level profiles (from 5 to 6). Table 2 shows the interestingness parameter values for the six rules.

Multi-risk profiles allow us to better characterize workers' sensibility to risk. Indeed, we are able to comprehend which factors act as a trigger in the behavioural strategies for risk management. For instance, as rule 3 shows, a high level of expertise favours a very careful behaviour: a worker classified in this profile has no need of sensibilization to risk. On the other hand, we have found that a worker with low work gratification who prevents falling by only avoiding climbing, potentially performs no actions to prevent intoxication events (rule 6).

As a final remark, we wish to observe that, in both the previous cases, the choice for the minimum support, confidence and Kulc index stems from the desire to find out the true relationship between criticality factors and risk-related behaviours. In order to achieve useful association rules, the minimum support should be neither too low (not to generate a huge number of uninteresting rules) nor too high (not to produce too few rules expressing common sense knowledge); further, the minimum thresholds for confidence and Kulc index should both be high, respectively, to provide evidence of the validity of the discovered relationship and to denote a strong positive correlation among the involved items. On the other hand, we should appropriately lower the minimum support threshold if we want to discover uncommon profiles that represent interesting outliers for the specific application.

5. Conclusions

In this paper, we have presented an association-rule based approach to workers' profiling based on their sensibility degree to the risks they are exposed to in the workplace. The developed profiler has been tested on a small dataset consisting of 80 transactions (i.e., interviews with as many workers) containing risks, prevention actions and criticality factors. The strong association rules, i.e., the rules associated with interestingness measures (namely, support, confidence and Kulc index) higher than predefined thresholds, represent risk sensibility profiles. In particular, a strong single-risk rule can be interpreted as a statement about the relationship between specific values of the criticality factors and the actions that are able to prevent the considered risk. On the other hand, a strong multi-risk rule may represent two different concepts. The former is simply a connection between given values of criticality factors and the actions preventing the occurrence of the set of risks taken into account. The latter relates the set of given values of criticality factors and specific actions aimed at preventing the occurrence of a subset of the involved risks, with the actions performed to prevent the remaining considered risks.

The main novelty of the developed risk sensibility profiler is that it represents a valuable objective tool to find out possible interrelations among risk perception, risk propensity and risk-related decision-making. By automatically discovering the typical ways in which workers perceive risks and would react to them, a risk management trainer can profitably adopt a training process specifically tailored to each worker's needs.

References

Agrawal, R., Imieliński, T., & Swami, A.(1993a). Database mining: A performance perspective. *IEEE Transactions on Knowledge and Data Engineering*, *5*, 914–25.

Agrawal, R., Imieliński, T., & Swami, A.(1993b). Mining association rules between sets of items in large databases. In ACM(Ed.), *Proc. ACM SIGMOD Int. Conf. on Management of Data* (pp. 207–16). Washington, D.C.

Agrawal, R., & Srikant, R.(1994). Fast algorithms for mining association rules. *Proceedings of the 20th VLDB Conference*, (pp. 487–99). Santiago, Chile.

Aven, T.(2012). Foundational issues in risk assessment and risk management. *Risk Analysis*, . Doi: 10.1111/j.1539-6924.2012.01798.x.

Bouyer, M., Bagdassarian, S., Chaabanne, S., & Mullet, E.(2001). Personality correlates of risk perception. *Risk Analysis*, *21*, 457–65.

Bradshaw, J.(2001). Yams - yet another measure of similarity. In *EuroMUG, Cambridge, UK*. `http://www.daylight.com/meetings/emug01/Bradshaw/Similarity/YAMS.html`.

Chauvin, B., Hermand, D., & Mullet, E.(2007). Risk perception and personality facets. *Risk Analysis*, *27*, 171–85.

Costa-Font, J., Mossialos, E., & Rudisill, C.(2009). Optimism and the perception of new risks. *Journal of risk research*, *12*, 27–41.

Dubois, D.(2010). Representation, propagation, and decision issues in risk analysis under incomplete probabilistic information. *Risk Analysis*, *30*, 361–8.

Han, J., Cheng, H., Xin, D., & Yan, X.(2007). Frequent pattern mining: current status and future directions. *Data Mining and Knowledge Discovery*, *15*, 55–86.

Horswill, M. S., & McKenna, F. P.(1999). The effect of perceived control on risk taking. *Journal of Applied Social Psychology*, *29*, 377–91.

ISO27001(2005). *ISO 27001*. International Organization for Standardization.

ISO31000(2009). *ISO 31000 Risk management - Principles and guidelines*. International Organization for Standardization.

Keil, M., Wallace, L., Turk, D., Dixon-Randall, G., & Nulden, U.(2000). An investigation of risk perception and risk propensity on the decision to continue a software development project. *The Journal of Systems and Software*, *53*, 145–57.

Klein, C. T. F., & Helweg-Larsen, M.(2002). Perceived control and the optimistic bias: a meta-analytic review. *Psychology and health*, *17*, 437–46.

Kulczynski, S.(1927). Die panzenassoziationen der pieninen. *Bulletin International de l'Academie Polonaise des Sciences et des Lettres, Classe des Sciences Mathematiques et Naturelles*, (pp. 57–203).

Lazzerini, B., & Mkrtchyan, L.(2010). Risk analysis using extended fuzzy cognitive maps. In IEEE(Ed.), *Proc. Int. Conf. on Intelligent Computing and Cognitive Informatics* (pp. 179–82). Kuala Lumpur, Malaysia.

Lazzerini, B., & Mkrtchyan, L.(2011). Analyzing risk impact factors using extended fuzzy cognitive maps. *Systems Journal, IEEE*, *5*, 288–97.

Leitch, M.(2010). Iso 31000:2009. The new international standard on risk management. *Risk Analysis*, *30*, 887–92.

Misra, K. B.(2008). Risk analysis and management: An introduction. In K. B. Misra(Ed.), *Handbook of Performability Engineering* (pp. 667–81). Springer London.

Nicholson, N., Fenton-O'Creevy, M., Soane, E., & Willman, P.(2001). Risk propensity and personality. *Social Research*, *8*, 1–33.

Peters, E., & Slovic, P.(1996). The role of affect and worldviews as orienting dispositions in the perception and acceptance of nuclear power. *Journal of applied social psychology*, *26*, 1427–8.

Sharma, D., Alford, B., Bhuian, S., & Pelton, L.(2009). A higher-order model of risk propensity. *Journal of Business Research*, *62*, 741–4.

Sitkin, S., & Weingart, L.(1995). Determinants of risky decision-making behavior: a test of the mediating role of risk perception and propensity. *Academy of management*, *38*, 1573–92.

Sjöberg, L.(2000). Factors in risk perception. *Risk Analysis*, *20*.

Slovic, P.(1987). Perception of risk. *Science*, *236*, 280–5.

Sueiro, A., Sánchez-Iglesias, I., & de Tella, A. M.(2011). Evaluating risk propensity using an objective instrument. *Span J Psychol.*, *14*, 392–410.

Wu, T., Chen, Y., & Han, J.(2010). Re-examination of interestingness measures in pattern mining: a unified framework. *Data Mining and Knowledge Discovery*, *21*, 371–97.