Environment Systems and Decisions (2022) 42:436–449
https://doi.org/10.1007/s10669-022-09857-z

A Novel Metric to Evaluate the Association Rules for Identification of Functional Dependencies in Complex Technical Infrastructures

Federico Antonello[1] · Piero Baraldi[1] · Enrico Zio[1,2,3] · Luigi Serio[4]

Accepted: 28 March 2022 / Published online: 20 May 2022

Abstract

Functional dependencies in complex technical infrastructures can cause unexpected cascades of failures, with major consequences on availability. For this reason, they must be identified and managed. In recent works, the authors have proposed to use association rule mining for identifying functional dependencies in complex technical infrastructures from alarm data. For this, it is important to have adequate metrics for assessing the effectiveness of the association rules identifying the functional dependencies. This work demonstrates the limitations of traditional metrics, such as lift, interestingness, cosine and laplace, and proposes a novel metric to measure the level of dependency among groups of alarms. The proposed metric is compared to the traditional metrics with reference to a synthetic case study and, then, applied to a large-scale database of alarms collected from the complex technical infrastructure of CERN (European Organization for Nuclear Research). The results confirm the effectiveness of the proposed metric of evaluation of association rules in identifying functional dependencies.

Keywords Complex technical infrastructure · Functional dependency · Association rule mining · Alarm data

Abbreviations

CTI	Complex technical infrastructure
FDEP	Functional dependency
N_c	Number of CTI components
c_j	Generic component j
a_j^k	Alarm associated to the malfunction of type k of component j
M_j^{al}	Number of types of alarms triggered by component j
M^{al}	Total number of types of alarms
$A = \left\{a_j^k\right\}$	Set of all possible types of alarms
N^{al}	Total number of alarms collected in the database
$[t_0, t_f]$	Time domain during which the N^{al} alarm messages of the database have been collected
Z	Number of time intervals in which the time domain $[t_0, t_f]$ is subdivided
Δt	Time interval length
$s_j^k(z)$	Boolean variable associated to the occurrence of the alarm a_j^k in the z time interval z
$\vec{c}_j(z)$	Vector of size M_j^{al} indicating the state of the component j in the time interval z
$\vec{T}(z)$	Vector of size M^{al} indicating the state of the CTI in the time interval z
T	Matrix of size $[\mathrm{Z} \times M^{al}]$ representing the evolution of the CTI state in the time domain $[t_0, t_f]$
X	Generic pattern of alarms
$n(X)$	Number of time intervals in which at least all the alarms of X occur
$S(X)$	Support of X
$P(X)$	Probability of occurrence of X
X^{fp}	Generic frequent pattern of alarms
$s\%$	Minimum support threshold
$c\%$	Minimum confidence threshold
$r_l = \left\{x_l^a \Rightarrow x_l^a\right\}$	nGeneric association rule
x^a	Antecedent of the *l-th* association rule

✉ Piero Baraldi
piero.baraldi@polimi.it

1 Energy Department, Politecnico di Milano, Via Lambruschini 4, 20156 Milan, Italy

2 MINES ParisTech, PSL Research University, CRC, Sophia Antipolis, France

3 Eminent Scholar, Department of Nuclear Engineering, College of Engineering, Kyung Hee University, Seoul, Republic of Korea

4 Engineering Department, CERN, 1211 Geneva 23, Switzerland

y^a	Consequent of the *l-th* association rule
$Conf(x_l^a \Rightarrow x_l^a)$	Confidence of the *l-th* association rule
AR	Set of the obtained association rules
λ_j^k	Transition rate of component *j* out of the generic state *k*

1 Introduction

Functional Dependencies (FDEPs) play a crucial role in the operation of Complex Technical Infrastructures (CTIs) but can also constitute channels of vulnerability through which failures can cascade (Johansson and Hassel 2010; Eusgeld et al. 2011; Kröger and Zio, 2011; Zio, 2016a, b; Hickford et al. 2018; Azzolin et al. 2018; Huai et al. 2018; Cantelmi et al. 2021). In fact, local malfunctions or perturbations can propagate, due to FDEPs, through groups of dependent components, originating unexpected cascades of failures across systems, which can lead to large-scale consequences of CTI unavailability and event unexpected damages of equipment (Johansson and Hassel 2010; Eusgeld et al. 2011; Kröger and Zio, 2011; Moura et al. 2015; Zio 2016a, b; Antonello et al. 2019). Examples of CTIs which can be affected by FDEPs are electric power grids, natural gas pipelines, water distribution networks, transportation networks, large research facilities and internet communication networks (Zhang et al. 2016; Ballantyne et al. 2018; Li and Liu 2018). One evidence of cascading failures caused by FDEPs in CTIs is the power blackout that triggered a shortage of water supply and propagated by inducing malfunctions and failures in other critical infrastructures in the eastern USA on 14th August 2003. Similarly, in India, a major blackout triggered by a power relay failure led to water supply and oil supply interruption communications interruption, and serious traffic congestion, which paralyzed the region for a long period of time (Jin et al. 2017). Other examples include the 1998 Canada ice storm (Chang et al. 2007), the 2001 US World Trade Center attack (Mendonça & Wallace, 2006), the 2003 North American blackout (U.S.-Canada Power System Outage Task Force, 2004), the 2010 Chile earthquake and tsunami (Wen et al. 2010) and the power blackout resulted from Hurricane Sandy in 2012, which led to a loss of water supply in New York City.

Traditional methods for FDEPs identification typically require a deep knowledge of the systems, including its logic and structure function. In practice, this is not easy to retrieve for complex and evolving CTIs (Billinton and Allan, 1992; Xing et al. 2014). Recent works have explored the possibility of using data-driven methods applied to alarm messages collected by supervision systems (Serio et al. 2018; Antonello et al. 2019). An Association Rule Mining (ARM) method which scan alarm databases to identify FDEPs and groups of functionally dependent components has been proposed in (Antonello et al. 2021a). The method relies on an Apriori-based algorithm (Srikant and Agrawal 1996) that extracts patterns of alarms frequently occurring together and derives the association among them in the form of "*if* (*condition*) *then* (*consequence*)" rules. An association rule is considered only if the frequency of occurrence of the involved alarms is larger than a predefined threshold, called *minimum support*, and if the conditional probability of occurrence of the *consequence* given the *condition* is larger than a predefined threshold, called *minimum confidence* (Srikant and Agrawal, 1996; Hui et al. 2005). On the other hand, the identification of rare functional dependencies related to cascading malfunctions and perturbations for vulnerability analysis (Antonello et al. 2021b), requires using small values of *minimum support* with the consequences of *a)* finding many spurious rules, i.e. rules including alarms that do not belong to a real FDEP but are included by chance (Zhang et al. 2016; Hämäläinen and Webb 2019; Antonello et al. 2022) and *b)* finding a very large number of rules (Van Leeuwen and Galbrun, 2015; Marinica and Guillet, 2010). Therefore, the generated rules must be post-processed to identify the FDEPs of interest by experts of the system. This labour intensive task can, in principle, be alleviated by using metrics, such as *confidence*, *lift*, *interestingness*, *cosine* and *laplace,* that have been proposed to assess the strength of association rules (Benites and Sapozhnikova 2014; Luna et al. 2018). However, these metrics, which consider statistical properties of the data, such as the probability of occurrence of a rule in the database, the co-occurrence probability of the *condition* given the *consequence* part (and vice versa) and the probability of occurrence of the *condition* and/or the *consequence* (Luna et al. 2018), are not tailored on the user necessities (Mathu et al. 2011). In practice, they application to the identification of FDEPs do not guarantee that the interesting rules, such as those rare and without spurious alarms, will be extracted (Marinica and Guillet, 2010). Moreover, as suggested in (Mathu et al. 2011), rule post-processing should be based on user necessities.

In this context, the objective of this work is twofold: (1) analyse the effectiveness and make emerge the limitations of the traditional metrics used to evaluate the association rules identifying FDEPs; (2) propose a novel metric to effectively *i)* identify FDEPs from alarm data, but also *ii)* discriminate between spurious and actual FDEPs. The proposed metric is constructed considering the definition of conditional probability and the need of avoiding the inclusion of spurious alarms in functionally dependent groups.

The main contributions of the work are:

the analysis and the identification of criticalities in the post-processing approach based on metrics of association rule quality with respect to FDEPs identification;

the proposal of a metric for discovering groups of non-spurious, functionally dependent alarms.

The effectiveness of the proposed metric is analysed by comparison with the most commonly used metrics for association rules considering *i)* a synthetic case study and *ii)* a large-scale database of alarms generated by different supervision systems of a representative subset of the CTI of CERN.

The remainder of the paper is organized as follows: Sect. 2 describes the context of the work and introduces the association rules mining. In Sect. 3, the proposed metric is presented. Section 4 discusses the case studies and the obtained results. Finally, Sect. 5 draws some conclusions and recommends potential future lines of work.

2 Context of the work

2.1 Complex technical infrastructure

In this work, a CTI composed by N_c components is considered together with a database of a large number of alarm messages,$N^{al} >> 1$, generated during its operation over a long period of time $[t_0, t_f]$. The generic *i-th* alarm message is associated to the pair (t_i, m_i) of the time t_i at which the alarm of type m_i occurs. Assuming that there are M_j^{al} types of alarms associated to the generic *j-th* component, c_j, a label a_j^k is introduced for the *k-th* type of alarm message associated to component c_j. The set A containing all types of alarm messages in the database is:

$$A = \left\{ a_1^1, \dots, a_1^{M_1^{al}}, \dots, a_j^1, \dots a_j^{M_j^{al}}, \dots, a_{N_c}^1, \dots, a_{N_c}^{M_{N_c}^{al}} \right\} \tag{1}$$

and the total number of types of alarm messages is:

$$M^{al} = \sum_{j=1}^{N_c} M_j^{al} \tag{2}$$

Following (Antonello et al. 2019), the overall time interval $[t_0, t_f]$ during which the alarm messages have been collected is subdivided into Z consecutive small time intervals of the same length $\Delta t = \frac{t_f - t_0}{Z}$ and a Boolean variable, $s_j^k(z)$, $z = 1, \dots, Z$, is associated to the occurrence of alarm a_j^k in the *z-th* time interval:

$$s_j^k(z) \begin{cases} 1 \; if\ alarm\ a_j^k\ occurs\ at\ least\ once\ in [t_0 + (z-1) \cdot \Delta t, t_0 + z \cdot \Delta t) \\ 0 \; otherwise \end{cases} \tag{3}$$

The state of the CTI during the generic *z-th* time interval is represented by the Boolean vector:

$$\vec{\mathcal{T}}(z) = \left[s_{j'}^{k'}(z), \dots, s_{j''}^{k''}(z), \dots, s_{N_c}^{M_j^{al}}(z) \right] \in [0,1]^{M^{al}} \tag{4}$$

According to this notation, the database of alarms $(t_i, m_i), i = 1, \dots, N^{al}$, is represented by the Boolean matrix:

$$T = \begin{bmatrix} \vec{\mathcal{T}}(1) \\ \dots \\ \vec{\mathcal{T}}(Z) \end{bmatrix} \in [0,1]^{Z \times M^{al}} \tag{5}$$

whose generic *z-th* row refers to the state of the CTI during the *z-th* time interval. Therefore, T provides a dynamic representation of the CTI state evolution in the time interval $[t_0, t_f]$.

3 Association rules mining

Association rules mining algorithms have been used for identifying FDEPs in relational databases (Sánchez et al. 2008; Yao and Hamilton 2008) and have been tailored in (Antonello et al. 2019) to identify FDEPs from alarm data collected in CTIs. Considering a pattern (subset) of alarms $X \subseteq A$, an association rule is a probabilistic logical expression of the form $x^{\boldsymbol{a}} \Rightarrow y^{\boldsymbol{a}}, x^{\boldsymbol{a}} \subset X, y^{\boldsymbol{a}} = X - x^{\boldsymbol{a}}$, representing the conditional co-occurrence of the two subsets, $x^{\boldsymbol{a}}$ and $y^{\boldsymbol{a}}$, of the pattern $X \subseteq A$, where $x^{\boldsymbol{a}}$ and $y^{\boldsymbol{a}}$ are referred to as "antecedent" and "consequent" of the rule, respectively (Srikant and Agrawal, 1996; Hui et al. 2005). Let $n(X)$ be the counter of the number of vectors $\vec{\mathcal{T}}(z)$ of the database $T = \begin{bmatrix} \vec{\mathcal{T}}(1) \\ \dots \\ \vec{\mathcal{T}}(Z) \end{bmatrix}$ characterized by the occurrences of at least all the alarms of X (i.e., $\forall a_j^k \subset X, s_j^k(z) = 1$) and $S(X)$ be the *support* of X, which represents an estimation of its probability of occurrence, $P(X)$:

$$S(X) = \frac{n(x^{\boldsymbol{a}} \cup y^{\boldsymbol{a}})}{Z} \tag{6}$$

The expression $x^{\boldsymbol{a}} \Rightarrow y^{\boldsymbol{a}}$ is an association rule if:

a) the *support* of X is larger than a minimum support threshold *ms*:

$$S(X) = S(x^{\boldsymbol{a}} \Rightarrow y^{\boldsymbol{a}}) > ms \tag{7}$$

b) the *confidence* of the rule $x^a \Rightarrow y^a$ is larger than a minimum confidence threshold *mc*:

$$Conf(x^a \Rightarrow y^a) = \frac{S(X)}{S(x^a)} > mc \tag{8}$$

where the *confidence* is an estimate of the conditional probability of occurrence $P(y^a|x^a)$ of the subset y^a given the occurrence of x^a.

In this work, the Apriori algorithm is used for association rules mining (Srikant and Agrawal 1996; Zaki 2000; Witten and Frank 2016; Antonello et al. 2021a). It is based on the two steps of:

Identification of frequent subsets of alarms, $X^{fp} \subseteq A$ characterized by a support larger than ms, i.e., $S(X^{fp}) > ms$;

Extraction of association rules $x^a \Rightarrow y^a, x^a \subset X^{fp}$, $y^a = X^{fp} - x^a$ from the frequent subsets of alarms identified in 1). An association rule should satisfy the confidence condition $C(x^a \Rightarrow y^a) > mc$.

4 Association rules metrics

The effectiveness of an association rule, $x^a \Rightarrow y^a, x^a \subset X$, $y^a = -x^a$, is typically evaluated considering ad-hoc defined metrics, such as the *lift* (Luna et al. 2018), *interestingness* (Ghosh and Nath, 2004), *cosine* (Luna et al. 2018) and *laplace* (Geng and Hamilton, 2006).

The *lift* metric is defined as:

$$Lift(x^a \Rightarrow y^a) = \frac{P(X)}{P(x^a) \cdot P(y^a)} \tag{9}$$

The numerator is the probability of the joint occurrence of the antecedent and the consequent ($P(X)$) and the denominator represents the same quantity, but under the hypothesis of independence between antecedent and consequent (i.e., $P(x^a) \cdot P(y^a)$). Then, the *lift* measures the mutual dependency among the alarms in the rule antecedent and consequent parts (Luna et al. 2018). The larger the *lift*, the stronger the mutual dependency among x^a and y^a. In particular, values of *lift* equal to (lower than) 1, indicate that the rule antecedent and consequent parts are uncorrelated (negatively correlated). The value of *lift* can be estimated from a dataset of alarms as:

$$\widehat{Lift}(x^a \Rightarrow y^a) = \frac{S(X)}{S(x^a) \cdot S(y^a)} \tag{10}$$

Cosine is a metric derived from the *lift* (Luna et al. 2018):

$$Cosine(x^a \Rightarrow y^a) = \frac{P(X)}{\sqrt{P(x^a) \cdot P(y^a)}} = \sqrt{Lift(x^a \Rightarrow y^a) \cdot P(X)} \tag{11}$$

whose main difference with *lift* and *cosine* is that its estimate:

$$\widehat{Cosine}(x^a \Rightarrow y^a) = \frac{S(X)}{\sqrt{S(x^a) \cdot S(y^a)}} = \sqrt{Lift(x^a \Rightarrow y^a) \cdot S(X)} \tag{12}$$

is not proportional to the total number of time intervals Z in the database (Luna et al. 2018).

Another metric proposed to evaluate the strength of a rule is the *Interestingness* (Ghosh and Nath, 2004):

$$\begin{aligned} Interestingness(x^a \Rightarrow y^a) &= \frac{P(X)}{P(x^a)} \\ &\cdot \frac{P(x)}{P(y^a)} \cdot (1 - P(X)) = Conf(x^a \Rightarrow y^a) \\ &\cdot Conf(y^a \Rightarrow x^a) \cdot (1 - P(X)) \end{aligned} \tag{13}$$

which is proportional to the following three quantities: *1)* how much the consequent is dependent on the antecedent (i.e., $Conf(x^a \Rightarrow y^a)$), *2)* how much the antecedent is dependent on the consequent (i.e., $Conf(y^a \Rightarrow x^a)$), and *3)* how rare is the rule (i.e., $(1 - P(X))$). According to this metric, the most interesting rules are characterized by a strong mutual dependency between antecedent and consequent and are rare.

Finally, we consider the *Laplace* metric, which is a metric specifically derived from confidence to account for statistical fluctuations when the support is estimated from databases:

$$\widehat{Laplace}(x^a \Rightarrow y^a) = \frac{S(X) \cdot Z + 1}{S(x^a) \cdot Z + 2} = \frac{Z \cdot S(X) + 1}{Z \cdot \frac{S(X)}{Conf(x^a \Rightarrow y^a)} + 2} \tag{14}$$

This metric is monotone in both *support* and *confidence*, and has been proven to be useful to identify only rules with large *support* and *confidence* at the same time (Geng and Hamilton, 2006).

Table 1 reports the main metrics proposed for evaluating association rules, considering both their probabilistic formulation and their estimation from a database.

Table 1 Metrics for association rules evaluation

Probabilistic formulation	Estimation from a database
$Lift = \frac{P(X)}{P(x^a) \cdot P(y^a)}$	$\widehat{Lift} = \frac{S(X)}{S(x^a) \cdot S(y^a)}$
$Interestingness = \frac{P(X)}{P(x^a)} \cdot \frac{P(X)}{P(y^a)} \cdot (1 - P(X))$	$\widehat{Interestingness} = \frac{S(X)}{S(x^a)} \cdot \frac{S(X)}{S(y^a)} \cdot (1 - S(X))$
$Cosine = \frac{P(X)}{\sqrt{P(x^a) \cdot P(y^a)}}$	$\widehat{Cosine} = \frac{S(X)}{\sqrt{S(x^a) \cdot S(y^a)}}$
Not Defined	$\widehat{Laplace} = \frac{n(X)+1}{n(x^a)+2}$

5 A novel metric for evaluating association rules identifying FDEPs

5.1 Probabilistic description of FDEPs

According to (Etesami and Kiyavash 2017), two components of a system are functionally dependent if the operation of one is influenced by the operation of the other. Notice that functionally dependent components are usually causally related and FDEPs are unidirectional, e.g., a compressor requires the functioning of the electrical systems to work properly, whereas the electrical system do not require the functioning of the compressor to operate. However, although the occurrences of malfunctions and failures triggered by FDPEs are temporally ordered, e.g. the malfunction of the electric system precedes the malfunction of the compressor, several methods for the identification of FDEPs, such as those based on the extraction of association rules from alarms (Serio et al. 2018; Antonello et al. 2020) or traditional parametric models for root cause of failures analysis, such as beta factor model, binomial failure rate model etc. (Mosleh, 1991; Zio, 2009; O'Connor and Mosleh, 2016), do not consider the temporal order of the chains of events involved in the FDEPs. Similarly in this work we focus only on the identification of the group of functionally dependent components involved in the FDEPs, which can be successively analysed by operators and experts to reconstruct the causal chain of events, or processed by ad-hoc algorithms tailored to identify causal relationships from groups of alarms (Antonello et al. 2020).

Considering alarm messages triggered when components have abnormal behaviours or non-nominal performances for the relevant operating mode, two generic components of a CTI, c_1 and c_2, are assumed to be functionally dependent if an abnormal behaviour of component c_1, revealed by an alarm,a_1^k, increases the probability of occurrence of a malfunction of components c_2, revealed by another alarm,a_2^k, or viceversa.

From the probabilistic point of view, the two malfunctions revealed by the alarms a_1^k and a_2^k, whose probabilities of occurrence are $P(a_1^k)$ and $P(a_2^k)$, are functionally dependent if the probability of their co-occurrence, $P(a_1^k \cap a_2^k)$, is:

$$P(a_1^k \cap a_2^k) > P(a_1^k) \cdot P(a_2^k) \tag{15}$$

In general, considering a pattern of R alarms $X_{FDEP} = \left\{ a_{j^1}^{k^1}, \dots, a_{j^R}^{k^R} \right\}$ triggered by a functional dependency, its probability of occurrence is:

$$P\left(X_{FDEP}\right) > P(a_{j^1}^{k^1}) \cdot \dots \cdot P(a_{j^R}^{k^R}) \tag{16}$$

Therefore, in case of functional dependency among the alarms of X_{FDEP}, the ratio:

$$I_{X_{FDEP}} = \frac{P\left(X_{FDEP}\right)}{\prod_{r=1}^{R} P(a_{j^r}^{k^r})} \tag{17}$$

is expected to be larger than 1. In particular, the stronger the dependency among the alarms of the pattern X_{FDEP}, the larger the ratio $I_{X_{FDEP}}$. The index $I_{X_{FDEP}}$ is defined as the *lift* of the itemset X_{FDEP} and used as metric to assess dependencies among the items of an association rule. If the probability of occurrence of a generic pattern $X = \left\{ a_{j^1}^{k^1}, \dots, a_{j^R}^{k^R} \right\}$ is estimated from the alarm database using its support, Eq. 15 becomes:

$$\hat{I}_X = \frac{S(X)}{\prod_{r=1}^{R} S(a_{j^r}^{k^r})} \tag{18}$$

6 A novel metric

A drawback of the index $\hat{I}_X$ is that it does not allow distinguishing the presence of a spurious (independent) alarm within a group of functionally dependent alarms. Let us consider a pattern X_S made by R functionally dependent alarms $X_{FDEP} = \left\{ a_{j^1}^{k^1}, \dots, a_{j^R}^{k^R} \right\}$ and a spurious alarm a_{js}^{ks}, i.e. $X_S = X_{FDEP} \cup a_{js}^{ks}$. By applying Eq. 15, the index I_{X_S} is:

$$\begin{aligned} I_{X_S} &= \frac{P\left(X_S\right)}{\prod_{r=1}^{R} P(a_{j^r}^{k^r}) \cdot P(a_{js}^{ks})} = \frac{P\left(X_{FDEP}\right) \cdot P(a_{js}^{ks})}{\prod_{r=1}^{R} P(a_{j^r}^{k^r}) \cdot P(a_{js}^{ks})} \\ &= \frac{P\left(X_{FDEP}\right)}{\prod_{r=1}^{R} P(a_{j^r}^{k^r})} = I_{X_{FDEP}} \end{aligned} \tag{19}$$

Considering the statistical fluctuation in the occurrence of the alarms, it may occur that $\hat{I}_{X_{FDEP}} < \hat{I}_{X_S}$ and, therefore, the index $\hat{I}_X$ can provide misleading information on the FDEPs.

Notice that in case of FDEP among a pattern of alarms $X_{FDEP} = \left\{ a_{j^1}^{k^1}, \dots, a_{j^R}^{k^R} \right\}$, the probability of occurrence of each malfunction $P(a_{j^r}^{k^r})$, $r = 1, \dots, R$, can be decomposed into:

$$P(a_j^k) = P_{X_{FDEP}}(a_j^k) + P_{Ind}(a_j^k) \tag{20}$$

where $P_{X_{FDEP}}(a_j^k)$ and $P_{Ind}(a_j^k)$ are the probabilities that a_j^k occurs due to the FDEP and due to an event independent of the FDEP, respectively (Mosleh, 1991; Zio, 2009; O'Connor and Mosleh, 2016). In this work, where we are interested in rare FDEPs, we assume that the probability of occurrence $P(X_{FDEP})$ is greater than a fraction $\alpha \in [0, 1]$ of the total probability of occurrence , $P(a_j^k)$ of each alarma_j^k:

$$P(X_{FDEP}) > \alpha \cdot P(a_j^k), \forall a_j^k \in X \tag{21}$$

This assumption is motivated by: *a)* the probability that a generic alarm a_j^k occurs due to the rare FDEP is typically very close to the probability of occurrence of the whole pattern involved in the FDEP X_{FDEP}, $P_{X_{FDEP}}(a_j^k) \approx P(X_{FDEP})$, and *b)* to make possible the identification of the FDEP, its probability of occurrence $P(X_{FDEP})$ should be larger than the probability of the co-occurrence (by chance) of the generic alarm a_j^k and a spurious alarm a_{js}^{ks}, $P(X_{FDEP}) > P(a_j^k) \cdot P(a_{js}^{ks})$.

Consequently, if we consider a spurious alarm, a_{js}^{ks} and a rare FDEP $X_{FDEP} = \left\{a_{j^1}^{k^1}, \ldots, a_{j^R}^{k^R}\right\}$ with $P(X_{FDEP}) < \alpha$, the above mentioned assumption are not valid and, therefore, the co-occurrence probability,$P(X_S), X_S = X_{FDEP} \cup a_{js}^{ks}$, does not satisfy Eq. 21:

$$P(X_S) = P(X_{FDEP}) \cdot P\left(a_{js}^{ks}\right) < \alpha \cdot P\left(a_{js}^{ks}\right) \tag{22}$$

In this light, we define a novel metric for the evaluation of association rules identifying FDEPs:

$$I_{FDEP}(X) = \begin{cases} \frac{P(X)}{\prod_{a_j^k \in X} P(a_j^k)} \; if \;\; P(X) > \alpha \cdot max_{a_j^k \in X}\left(P\left(a_j^k\right)\right) \\ 0 \; otherwise \end{cases} \tag{23}$$

where $max_{a_j^k \in X}\left(P\left(a_j^k\right)\right)$ is the largest support among the alarms of X. Notice that given a pattern of alarm X, if the alarm with the largest support in X, $max_{a_j^k \in X}\left(P\left(a_j^k\right)\right)$, satisfies Eq. 21, also the other alarms satisfy it. In particular, given a generic rule $r_{FDEP} : \{x^a{}_{FDEP} \rightarrow y^a{}_{FDEP}\}$ which describes a real functional dependency involving the pattern of alarms $X_{FDEP} = x^a{}_{FDEP} \cup y^a{}_{FDEP}$, the metric $I_{FDEP}(X_{FDEP})$ is larger than 1. On the other hand, if we consider a spurious rule, $r_S : \{x^a{}_S T \rightarrow y^a{}_S\}$, $X_S = x^a{}_S \cup y^a{}_S$, including the alarm set X_{FDEP} and another alarm a_{js}^{ks} that does not belong to the actual functional dependency, i.e. $X_S = X_{FDEP} \cup a_{js}^{ks}$, $\alpha * max_{a_j^k \in X_S}\left(P\left(a_j^k\right)\right)$ is larger than the probability of occurrence of the pattern X_S and, therefore, the value of the metric $I_{FDEP}(X_S)$ is equal to 0.

Then, similarly to the metrics presented in Sects. 3.1 and 2.2 for which the probability of occurrence of a generic pattern X is estimated from the alarm database using its support, Eq. 23 becomes:

$$\hat{I}_{FDEP}(X) = \begin{cases} \frac{S(X)}{\prod_{a_j^k \in X} S(a_j^k)} if \;\; S(X) > \alpha \cdot max_{a_j^k \in X}\left(S\left(a_j^k\right)\right) \\ 0 \; otherwise \end{cases} \tag{24}$$

The proper setting of the parameter α should consider that, in practice, frequent functional dependencies are typically already known, whereas the rare ones are more likely to be unknown and relevant for CTI vulnerability (Lee et al. 2005; Antonello et al. 2019). Thus, considering Eq. 24, the use of a large value for α would drive the search to discover mainly the frequent FDEPs, with the risk of not identifying the rare FDEPs. On the opposite, some spurious patterns could be identified as actual FDEPs using very small values of α. For example, given a functionally dependent group of alarms X_{FDEP} with support $S(X_{FDEP}) = 0.01$ and a spurious alarm a_{js}^{ks} with support $S\left(a_{js}^{ks}\right) = 0.05$, their co-occurrence probability $P(X_S) = P(X_{FDEP}) \cdot P\left(a_{js}^{ks}\right)$, satisfies Eq. 20 when α is lower than 0.001, i.e. $P(X_{FDEP}) \cdot P\left(a_{js}^{ks}\right)$= $0.005 > \alpha \cdot P\left(a_{js}^{ks}\right)$.

7 Analysis of the effectiveness of the proposed metric in quantifying the interestingness of a rule

The effectiveness of a metric in quantifying the interest of a rule is typically verified considering various properties (Piatetsky-Shapiro 1991; Luna et al. 2018). In particular, (Piatetsky-Shapiro 1991) suggested that any quality metric M defined to quantify the interest of an association rule should be able to separate strong rules from weak ones by assigning larger values to the former. Also, M should verify the following three specific properties:

$M(x^a \Rightarrow y^a)$ of an association rule $x^a \Rightarrow y^a$ extracted from the database by chance should be 0;

$M(x^a \Rightarrow y^a)$ should monotonically increase with $P(X)$ when $P(x^a)$ and $P(y^a)$ remain constant. This guarantees that the larger the correlation among the rule antecedent x^a and consequent y^a, the more interesting the rule $x^a \Rightarrow y^a$;

$M(x^a \Rightarrow y^a)$ should monotonically decrease when $P(x^a)\,(P(y^a))$ increases if $P(X)$ andP$(y^a)\,(P(x^a))$ remain unchanged. This guarantees that the rarer is the antecedent, x^a, (or consequent,\,y^a), the more interesting is the rule.

Considering the metric I_{FDEP}, with respect to an association rule $x^a \Rightarrow y^a$, $y^a \cup x^a = X$, involving the alarms $a_j^k \in X$, property 1 is satisfied since $I_{FDEP}(x^a \Rightarrow y^a) \neq 0$ only if all the alarms $a_j^k \in X$ are functionally dependent. With respect to property 2, when the denominator of the expression of $I_{FDEP}(x^a \Rightarrow y^a)$, $\prod_{a_j^k \in X} P(a_j^k)$, remain constant, $P(x^a)$ and $P(y^a)$ remain constant, and, therefore, the larger $P(X)$, the larger $I_{FDEP}(x^a \Rightarrow y^a)$. Similarly, for what concerns property 3, when $P(X)$ is fixed, $I_{FDEP}(x^a \Rightarrow y^a)$ is proportional to $1/P(a_j^k)$, and therefore, when $P(x^a)$ (or $P(y^a)$) decreases, $\prod_{a_j^k \in x^a} P(a_j^k)$ decreases, and $I_{FDEP}(x^a \Rightarrow y^a)$ increases.

7.1 Case studies

The effectiveness of the traditional association rules metrics of *lift*, *interestingness*, *cosine*, *laplace* and of the proposed metric I_{FDEP} is evaluated considering *i)* a simulated system, and *ii)* a large-scale alarm database collected on a representative portion of CERN's CTI during 2016 (Serio et al. 2018; Antonello et al. 2019). The first application shows the limitations of the traditional metrics for evaluating association rules mining for FDEPs identification, the effectiveness of the proposed metric I_{FDEP} in discriminating spurious rules and the robustness of the results with respect to statistical fluctuations in the alarm database. The second application considers a real CTI, for which we expect to discover unknown and rare functional dependencies are not known a priori.

7.2 Simulated system

We consider a system made by 4 components $c_j, j = 1, 2, 3, 4$, which can be in the healthy, $k = 1$, or failed,$k = 2$, states and perform independent transitions at random times with constant failure rates. An alarm a_j^1 is triggered of each time component c_j performs a state transition from state 1 to state 2. The repair, i.e. the transition from the failed, $k = 2$, to the healthy, $k = 1$, state is assumed to be instantaneous. Table 2 reports the failure rates, from state 1 to state 2, of each component c_j.

We further model a functional dependency among components c_1, c_2, c_3. It is originated by the transition from state 1 to state 2 of a component c_3 which causes the instantaneous transition of components c_2 from state 1 to state 2. Then, the functional dependency can propagate to component c_1 by causing its transition from state 1 to state 2. The probability of the malfunction propagation, i.e., the probability that a transition of components c_2 from state 1 to state 2 causes the same state transition of component c_1, is set to 0.75.

Table 3 reports the probability of occurrence of all the possible patterns of alarms, $\subseteq \{a_1^1, a_2^1, a_3^1, a_4^1\}$, in 1 *a.t.u.*

Table 4 reports the values of the $I_X(X)$ and $I_{FDEP}(X)$ metrics for all the patterns of alarms, computed by applying Eqs. 17 and 23, respectively, when the parameter α of Eqs. 23 and 24 is set equal to 0.03. Notice that, as expected, the value of the metric $I_{FDEP}(X)$ is equal to 0 for all the spurious patterns of alarms (i.e., the patterns containing the spurious alarm a_4^1), and that the largest $I_{FDEP}(X)$ is assigned to the pattern $[a_2^1, a_3^1, a_1^1]$, which contains all the alarms involved in the FDEP. Since the Apriori-based algorithm tends to identify rules involving all the patterns (subsets) of alarms X' contained in the pattern (set) X_{FDEP} formed by all the alarms involved in the FDEP, the largest value of the proposed metric is correctly assigned to the whole set of dependent alarms (X_{FDEP}).

Table 2 Component failure rates in arbitrary time units (a.t.u.)$^{-1}$

Component c_j	*Failure rates* [a.t.u.$^{-1}$]
c_1	0.003
c_2	0.001
c_3	0.004
c_4	0.5

Table 3 Probabilities of occurrence of all the possible patterns of alarms

Subset of alarms	Occurrence probability
$[a_1^1]$	0.0060
$[a_2^1]$	0.0050
$[a_3^1]$	0.0040
$[a_4^1]$	0.5000
$[a_1^1, a_4^1]$	0.0030
$[a_1^1, a_2^1]$	0.0030
$[a_1^1, a_3^1]$	0.0030
$[a_2^1, a_3^1]$	0.0040
$[a_2^1, a_4^1]$	0.0025
$[a_3^1, a_4^1]$	0.0020
$X_{FDEP} = [a_1^1, a_2^1, a_3^1]$	0.0030
$[a_1^1, a_3^1, a_4^1]$	0.0015
$[a_2^1, a_3^1, a_4^1]$	0.0020
$[a_1^1, a_2^1, a_4^1]$	0.0015
$[a_1^1, a_2^1, a_3^1, a_4^1]$	0.0015

Table 4 Values of the $I_X(X)$ and $I_{FDEP}(X)$ metrics for all the subset of alarms of Table 5

X	$I_X(X)$	$I_{FDEP}(X)$
$[a_1^1, a_4^1]$	1	0
$[a_2^1, a_4^1]$	1	0
$[a_3^1, a_4^1]$	1	0
$[a_2^1, a_3^1]$	200	200
$[a_2^1, a_1^1]$	125	125
$[a_3^1, a_1^1]$	166	166
$X_{FDEP} = [a_2^1, a_3^1, a_1^1]$	24,999	24,999
$[a_2^1, a_3^1, a_4^1]$	200	0
$[a_2^1, a_1^1, a_4^1]$	125	0
$[a_3^1, a_1^1, a_4^1]$	166	0
$[a_2^1, a_3^1, a_1^1, a_4^1]$	24,999	0

The values of the metrics *lift*, *interestingness* and *cosine* have been computed by using the equations reported in the first column of Table 1 for all the 56 association rules that can be generated from the 15 patterns of Table 3. Table 5 reports the 8 association rules with the largest *lift*. Notice that rules N° 3 4 are also characterized by the largest *interestingness* and *cosine*. Among them, 6 rules are spurious (rules N° 1, 2, 5, 6, 7, 8) and only 4 rules are describing the actual functional dependency (rules N° 3, 4, 9, 10). Notice that, the spurious rules 1 and 2 are characterized by the largest value of *lift* (250) and their values of i*nterestingness*, *cosine* and *laplace* are larger than the values of rules 9, 10, which describe the actual functional dependency. Consequently, *lift*, *interestingness* and *cosine* cannot clearly discriminate spurious rules.

Then, to verify the effectiveness of the proposed metric $\widehat{I}_{FDEP}(X)$ computed from an alarm database, the behaviour of the 4 components-system has been simulated for a period of time $[t_0, t_f] = [0, 1500\ a.t.u.]$ and a database of 784 alarm messages has been obtained. The Boolean matrix, T, is computed following the procedure of Sect. 2.2 considering Z = 1500 time intervals of 1 *a.t.u.* and the metrics have been computed for all the 56 association rules that can be generated from the database. Table 6 reports the same list of association rules reported in Table 6, evaluated considering the same metrics and the *laplace* metric. The obtained results confirm that only the proposed metric $\widehat{I}_{FDEP}(X)$ is able to discriminate the spurious rules from a database of alarms.

To further analyze the robustness of the metrics with respect to the statistical fluctuations in the alarms occurrences, $N = 1000$ different alarm databases have been simulated and, for each database, the metrics have been computed. Table 7 reports the fraction of the databases in which the largest value of the considered metric is assigned to a spurious rule. As expected, only the value of the proposed metric, $\widehat{I}_{FDEP}$, is always larger for rules describing an actual functional dependency, whereas the other metrics are remarkably less robust to the statistical fluctuations.

This result is confirmed by Table 8 which reports the minimum, maximum and average values, over the 1000 alarm databases, of the metrics for rule 3, which refers to a real functional dependency, and for rule 7, which contains a spurious alarm. Notice that the $\widehat{I}_{FDEP}$ value of the spurious rule 3 is equal to zero in all the simulations, making clear its identification, whereas the other metrics tend to remarkably vary and they do not clearly discriminate the spurious rule from rule 7.

To verify the robustness of the proposed metric with respect to the probability of occurrence of the spurious alarm a_4^1, $P(a_4^1)$, has been varied in the range [0.005, 0.7],

Table 5 Association rules characterized by the largest values of lift, interestingness and cosine

N°	*Association rule*	*Lift*	*Interestingness*	*Cosine*	$I_{FDEP}(X)$
1	$[a_2^1, a_4^1] \rightarrow [a_3^1]$	250	0.49	0.71	0
2	$[a_3^1] \rightarrow [a_2^1, a_4^1]$	250	0.49	0.71	0
3	$[a_1^1, a_2^1] \rightarrow [a_3^1]$	250	0.75	0.87	24,999
4	$[a_3^1] \rightarrow [a_1^1, a_2^1]$	250	0.75	0.87	24,999
5	$[a_1^1, a_4^1] \rightarrow [a_3^1]$	250	0.37	0.61	0
6	$[a_3^1] \rightarrow [a_1^1, a_4^1]$	250	0.37	0.61	0
7	$[a_3^1, a_4^1] \rightarrow [a_1^1, a_2^1]$	250	0.37	0.61	0
8	$[a_2^1, a_4^1] \rightarrow [a_1^1, a_3^1]$	250	0.37	0.61	0
9	$[a_2^1, a_3^1] \rightarrow [a_1^1]$	125	0.38	0.61	24,999
10	$[a_1^1] \rightarrow [a_2^1, a_3^1]$	125	0.38	0.61	24,999

Table 7 Fraction of databases in which the largest value of the metric is assigned to a spurious rule

fraction of alarm databases in which the metric of a spurious rule is the largest, for each metric					
$\widehat{Lift}$	$\widehat{Interestingness}$	$\widehat{Cosine}$	$\widehat{Laplace}$	$\widehat{I}_X$	$\widehat{I}_{FDEP}$
0.858	0.230	0.230	0.174	0.481	0

Table 6 Estimation from the alarm database of the $\widehat{lift}$, $\widehat{interestingness}$, $\widehat{cosine}$, $\widehat{laplace}$, ratio $\widehat{I}_X$ and $\widehat{I}_{FDEP}$ for the rules identified in Table 5

N°	*Association rule*	$\widehat{Lift}$	$\widehat{Interestingness}$	$\widehat{Cosine}$	$\widehat{Laplace}$	$\widehat{I}_X$	$\widehat{I}_{FDEP}$
1	$[a_2^1, a_4^1] \rightarrow [a_3^1]$	71	0.14	0.38	0.50	74	0
2	$[a_3^1] \rightarrow [a_2^1, a_4^1]$	71	0.14	0.38	0.33	74	0
3	$[a_1^1, a_2^1] \rightarrow [a_3^1]$	142	0.43	0.65	0.80	6696	6696
4	$[a_3^1] \rightarrow [a_1^1, a_2^1]$	142	0.43	0.65	0.44	6696	6696
5	$[a_1^1, a_4^1] \rightarrow [a_3^1]$	57	0.11	0.34	0.42	74	*0*
6	$[a_3^1] \rightarrow [a_1^1, a_4^1]$	57	0.11	0.34	0.33	74	*0*
7	$[a_3^1, a_4^1] \rightarrow [a_1^1, a_2^1]$	333	0.66	0.82	0.75	9300	*0*
8	$[a_2^1, a_4^1] \rightarrow [a_1^1, a_3^1]$	333	0.66	0.82	0.60	9300	*0*
9	$[a_2^1, a_3^1] \rightarrow [a_1^1]$	93	0.28	0.53	0.66	6696	6696
10	$[a_1^1] \rightarrow [a_2^1, a_3^1]$	93	0.28	0.53	0.40	6696	6696

Table 8 Ranges of values of the metrics for rule 3 (containing all the alarms in the pattern X_{FDEP}= $[a_2^1, a_3^1, a_4^1]$) and the spurious rule 7

	Rule 3: $[a_1^1, a_2^1] \rightarrow [a_3^1]$			*Rule 7:* $[a_3^1, a_4^1] \rightarrow [a_1^1, a_2^1]$		
	Min value	*Max value*	*Average value*	*Min value*	*Max value*	*Average value*
$\widehat{Lift}$	68	750	191	42	1500	206
$\widehat{Interestingness}$	0.09	0.99	0.50	0.03	0.99	0.30
$\widehat{Cosine}$	0.30	1	0.70	0.20	1	0.50
$\widehat{Laplace}$	0.30	1	0.70	0.15	0.87	0.53
$\widehat{I}_X$	2066	187,499	17,643	1843	361,969	20,000
$\widehat{I}_{FDEP}$	2066	187,499	17,643	0	0	0

whereas the probabilities of occurrences of the other alarms and of the functional dependency have not been modified. Table 9 reports the fraction of the simulations in which a spurious rule obtains the largest value of the metrics over the $N = 1000$ database simulations, as a function of $P(a_4^1)$. Notice that, when the value of $P(a_4^1)$ is larger than 0.01 only the proposed metric correctly identifies the spurious rules, whereas when $P(a_4^1)$ is reduced to 0.01 or 0.005 the fraction of errors tend to increase. This is due to the fact that when $P(a_4^1)$ is lower than 0.05, the number of occurrences in the database of the spurious rules is equal to 1 or 2. Thus, when the support is very small the discriminator factor of Eq. 24 can be satisfied by chance. Notice, however, that the minimum support threshold of the Apriori algorithm prevents the generation of rules characterized by 1 or 2 occurrences. On the contrary, the $\widehat{lift}$, $\widehat{interestingness}$, $\widehat{cosine}$ and $\widehat{laplace}$ are likely to assign the largest value of the metric to a spurious rule with most of the values of $P(a_4^1)$. Notice that the smaller the fraction of simulations in which a spurious rule occurs, the smaller the probability of erroneously identifying a spurious rule as the one with the largest value of the metrics.

The sensitivity of the proposed metric with respect to the setting of the parameter α of Eq. 24 is also investigated. Figure 1 shows the fraction of simulations in which at least one spurious rule has the value of $\widehat{I}_{FDEP}$ larger than 0 and the fraction of simulations in which at least one rule containing a FDEP has $\widehat{I}_{FDEP}$ equal to 0, as a function of the parameter α. As discussed in Sect. 3.3, the lower the value of the parameter α, the larger the probability that a spurious group satisfies the discriminator factor of Eq. 24. In particular, when the parameter α is smaller than 0.01, the fraction of simulations in which the value $\widehat{I}_{FDEP}$ of spurious rules is larger than 0 tends to increase. On the opposite, the larger the value of the parameter α, the larger the probability that a rule containing a FDEP is not identified due to the condition $P(X) > \alpha \cdot max_{a_j^k \in X}\left(P\left(a_j^k\right)\right)$ in Eq. 24. Notice that when α is in the range [0.01, 0.08], the proposed metric allows correctly identifying all the FDEPs and discriminating all the spurious rules.

Finally, notice that the proposed metric, $\widehat{I}_{FDEP}$, depends only on the patterns $X = \left\{a_{j^1}^{k^1}, \ldots, a_{j^R}^{k^R}\right\}$ and do not require the definition of an association rule. This makes it possible to apply the metric also in the pattern mining domain, by evaluating the effectiveness of a pattern in describing a functional dependency, without the need of extracting associations rules, which is a time-consuming task (Del Jesus et al. 2011; Mukhopadhyay et al. 2014). Further, considering an alarm database consisting of M^{al} different alarm types, the number of possible patterns $\sum_{k=2}^{M^{al}} \frac{M^{al}!}{k!(M^{al}-k)!}$ is significantly smaller than the number of association rules $3^{M^{al}} - 2^{M^{al}+1} + 1$ that can be generated especially, as M^{al} increases (Del Jesus et al. 2011).

Table 9 Fraction of simulations in which the largest value of the metric is assigned to a spurious rule

$P(a_4^1)$	% of simulation in which at least a spurious rule occurs	Fraction of simulations in which the largest value of the metric is assigned to a spurious rule					
		$\widehat{Lift}$	$\widehat{Interestingness}$	$\widehat{Cosine}$	$\widehat{Laplace}$	$\widehat{I}_X$	$\widehat{I}_{FDEP}$
0.7	*99%*	0.901	0.399	0.399	0.346	0.505	0
0.5	95%	0.858	0.23	0.23	0.174	0.481	0
0.3	81%	0.729	0.102	0.102	0.076	0.467	0
0.1	46%	0.416	0.029	0.029	0.027	0.375	0
0.05	28%	0.249	0.016	0.016	0.012	0.215	0
0.01	6%	0.052	0.004	0.004	0.0	0.042	0.03
0.005	3%	0.026	0.001	0.001	0.0	0.022	0.02

Table 10 Example of the generated association rules

N°	Antecedent [System] {Alarm Identifier}⇒	Consequent {Alarm Identifier} [System]	Position in a ranking based on $\widehat{Lift}$	Position in a ranking based on $\widehat{Interestingness}$	Position in a ranking based on $\widehat{Laplace}$	Position in a ranking based on $\widehat{Cosine}$	Position in a ranking based on I_X	Position in a ranking based on I_{FDEP}
1	$\begin{bmatrix} Cryogenic \\ Cryogenic \\ CV \end{bmatrix}\begin{Bmatrix} a^4_{100} \\ a^4_{101} \\ a^2_{1123} \end{Bmatrix} \Rightarrow$	$\begin{Bmatrix} a^3_{123} \\ a^3_{124} \\ a^2_{102} \\ a^2_{1124} \end{Bmatrix}\begin{bmatrix} Cryogenic \\ Cryogenic \\ Cryogenic \\ CV \end{bmatrix}$	1	51	12	51	1	1
2	$\begin{bmatrix} Cryogenic \\ Cryogenic \\ Cryogenic \\ CV \end{bmatrix}\begin{Bmatrix} a^3_{123} \\ a^3_{124} \\ a^2_{102} \\ a^2_{1124} \end{Bmatrix} \Rightarrow$	$\begin{Bmatrix} a^4_{100} \\ a^4_{101} \\ a^2_{1123} \end{Bmatrix}\begin{bmatrix} Cryogenic \\ Cryogenic \\ CV \end{bmatrix}$	62	53	1	52	2	1
3	$\begin{Bmatrix} a^4_{100} \\ a^4_{102} \\ a^2_{101} \\ a^2_{1124} \end{Bmatrix}\begin{bmatrix} Cryogenic \\ Cryogenic \\ Cryogenic \\ CV \end{bmatrix} \Rightarrow$	$\{a^2_{1123}\}[CV]$	44	1	2	1	8	8
4	$[Electric]\{a^2_{5224}\} \Rightarrow$	$\{a^2_{4924}\}[Electric]$	98	3	26	5	12	12
5	$[CV]\{a^2_{1374}\} \Rightarrow$	$\{a^2_{1424}\}[CV]$	72	47	29	22	51	50

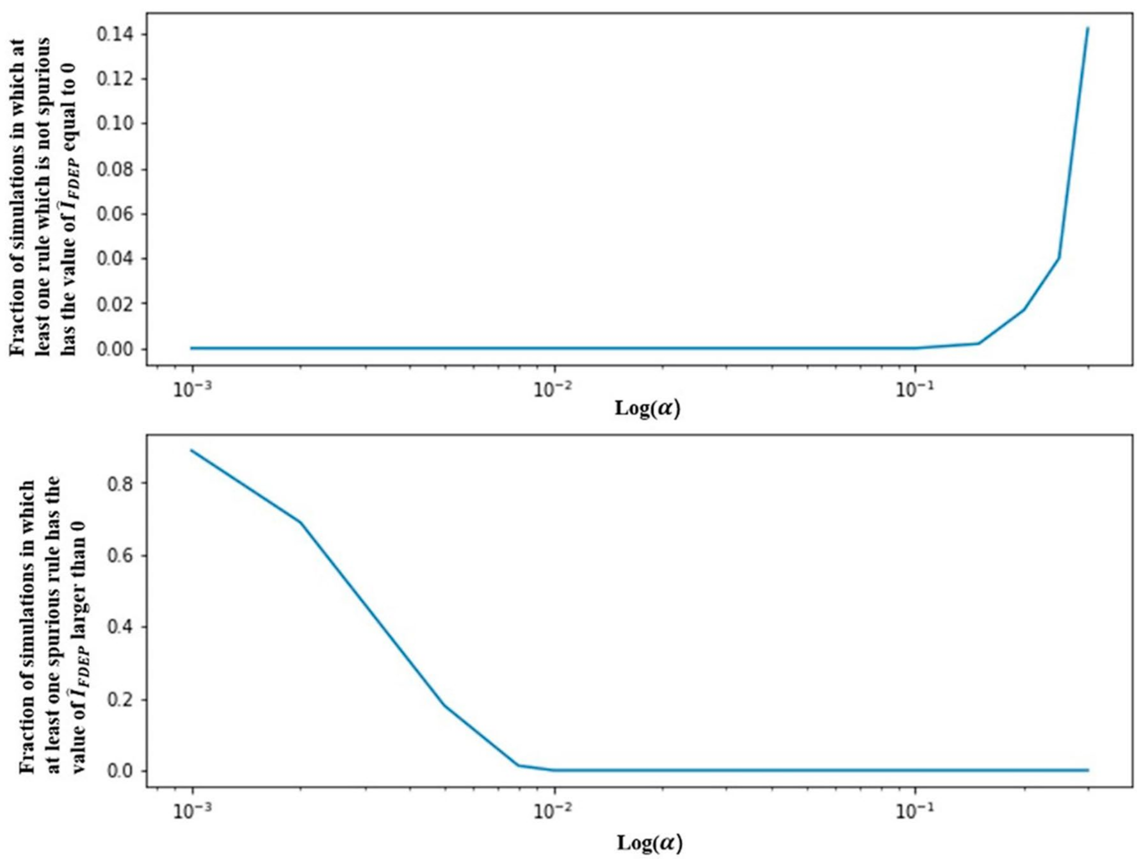

Fig. 1 Fraction of simulations in which at least one spurious rule has the value of $\widehat{I}_{FDEP}$ larger than 0 and the fraction of simulations in which at least one rule containing a FDEP has $\widehat{I}_{FDEP}$ equal to 0, as a function of the parameter α

7.3 CERN complex technical infrastructure

CERN (European Organization for Nuclear Research) LHC (Large Hadron Collider) is the largest existing particle accelerator and is made of a CTI composed by several systems working together for its functioning (Todd et al. 2016; Nielsen and Serio, 2016). It consists of a 27 km ring of superconducting magnets and engineered infrastructures, extending over the Swiss and French borders and located about 100 m underground. We consider the alarms databases generated during the period $[t_0, t_f]$ = [January 1st, 2016; December 31st, 2016] by the cryogenic, the electric and the cooling and ventilation systems in two adjacent sectors located in the*LHC point 8*, which is representative of the overall CTI complexity although composed only of ¼ of the overall CTI for these main three systems. A number N_{al} = 253,591 alarms reporting M^{al}=6800 different types of malfunctions caused by N_c = 2895 components have been collected during the considered period. The Apriori-based ARM approach proposed in (Antonello et al. 2019) has been applied considering 17,500 time intervals of length 30 min. Setting the *minimum support* and the *minimum confidence* thresholds equal to 0.02 and 0.8, respectively, 202 association rules have been obtained. Given the large number of obtained rules, it is important to prioritize them to identify the most interesting FDEPs of the CTI. To this aim, the rules have been ranked using the proposed metric $\widehat{I}_{FDEP}$ with α equal to 0.03, $\widehat{I}_X$ and the association rule metrics reported in Table 1.

Table 10 reports five rules with large $\widehat{I}_{FDEP}$ values and their position in the ranking based on the other metrics. Rules 1, 2 and 3 involve the same group of alarms and differ only for the combination in the antecedent and consequent part. This is consistent with the facts that the Apriori algorithm tends to identify rules involving all the subsets of the alarm generated from a FDEP, and that an association rule $r = \{x^a \Rightarrow y^a\}$ is a logical probabilistic expression and the rule direction,⇒, does not imply causality among the components in the antecedent and consequent part of the rule (Antonello et al. 2019). According to CERN experts, these rules, which are characterized by the largest value of $\widehat{I}_{FDEP}$, describe the propagation of a malfunction triggered by a problem in a cooling tower of the CV system (revealed by the alarms a^2_{1123}, a^2_{1124}) to the low pressure compressors (revealed by the alarms a^3_{100}, a^3_{101}, a^3_{102}), and the high pressure compressors of the Cryogenic system (revealed by the alarms a^4_{123}, a^4_{124}). The analysis of the major failure events occurred in the past has shown that this FDEP was part of chain of malfunctions responsible of a CTI shutdown occurred in 2016. Also, the same FDEP occurred in 33 other events during 2016, without causing the CTI shutdown. This is due to the fact that to fully propagate and lead the CTI shutdown, the cascade of events triggered by

Table 11 Example of spurious association rules

N°	Antecedent [System] {Alarm Identifier}⇒	Consequent {Alarm Identifier} [System]	Position in a ranking based on $\widehat{Lift}$	Position in a ranking based on $\widehat{Interestingness}$	Position in a ranking based on $\widehat{Laplace}$	Position in a ranking based on $\widehat{Cosine}$	Position in a ranking based on I_X	Position in a ranking based on I_{FDEP}
6	$\begin{bmatrix} CV \\ CV \end{bmatrix} \begin{Bmatrix} a^2_{1924} \\ a^2_{1374} \end{Bmatrix} \Rightarrow$	$\{a^2_{1424}\}[CV]$	85	166	9	41	16	198

Table 12 Description of the alarms involved in the rules of Tables 10 and 11

System	Alarm ID	Component ID	Component type	Alarm description
Cryogenic	a^3_{100}	c_{100}	Low pressure compressor	Water flow not ok
Cryogenic	a^3_{101}	c_{101}	Low pressure compressor	Water flow not ok
Cryogenic	a^3_{102}	c_{102}	Low pressure compressor	Water flow not ok
Cryogenic	a^4_{123}	c_{123}	High pressure compressor	Water flow not ok
Cryogenic	a^4_{124}	c_{124}	High pressure compressor	Water flow not ok
Cooling and Ventilation	a^2_{1123}	c_{1123}	Cooling Tower	Malfunction
Cooling and Ventilation	a^2_{1124}	c_{1124}	Cooling Tower	Malfunction
Cooling and Ventilation	a^2_{1924}	c_{1924}	Ventilator	Short circuit ventilation cycle
Cooling and Ventilation	a^2_{1374}	c_{1374}	Pump	Short circuit pump
Cooling and Ventilation	a^2_{1424}	c_{1374}	Pump	Short circuit pump

the FDEP require particular operating condition and/or the deterioration of other components involved in the operations. Therefore, its early identification would have made possible the implementation of preventive procedure to reduce its probability of occurrence and/or the impact of its consequences. In detail, the identified functionally dependent groups of alarms involved in a FDEP would have allowed performing detailed analyses on the monitored key physical quantities of the components associated to the alarms, to investigate the causes of the malfunction and prevent their future occurrences. This highlights the importance of the rules and, therefore, confirms the capability of the proposed metric $\widehat{I}_{FDEP}$ of identifying real functional dependencies among the CTI components. Similarly, the following rules of the $\widehat{I}_{FDEP}$ ranking describe relevant FDEPs occurred in the CTI of CERN.

Table 11 reports a rule containing, according to CERN experts, the spurious alarm a^2_{1924} generated by an independent component of the CV system, which is not related to the alarms a^2_{1374} and a^2_{1424}, generated by two functionally dependent pumps of a cooling circuit (see rules 5 of Table 10). Also, the actual functional dependence is captured in Rule 5 of Table 10, making the spurious rule unnecessary.

Notice that this rule is clearly identified as spurious only from the proposed metric $\widehat{I}_{FDEP}$, whereas the other metric values are large and would identify it as an actual rule.

Also, according to the CTI experts and engineers, the rules ranking obtained by the application of the proposed metric provides practical indications for:

updating the maintenance planning focusing on the importance of the discovered FDEPs; for example, increasing the frequency of the inspections of the component that causes the chain of events involved in a FDEP, with the objective of reducing the probability of their initiation;

discarding the spurious patterns of alarms, which can lead to possible errors of diagnosis of cascading failures;

implementing automatic or semi-automatic tools to support control room operators in real time in the management of alarms and failures. The discovered FDEPs can be used to warn the operators when an alarm involved in one of the discovered FDEPs occurs, in order to anticipate preventive interventions.

Table 12 provides the description of the alarms involved in the rules of Tables 10 and 11. The complete list of alarms involved in the database cannot be provided for confidentiality reasons.

8 Conclusions

A novel metric for the evaluation of association rules mined from alarm data to identify FDEPs has been proposed. The proposed metric allows ranking the association rules obtained from the ARM algorithm. This is a fundamental task in real applications where hundreds of rules are typically generated, the use of the traditional association rule metrics has been shown to be not effective and their one-by-one analysis is a time-consuming task. The effectiveness of the proposed metric has been shown by its application to the association rules generated from a simulated alarm database and a large-scale alarm database collected at CERN's CTI during 2016. The obtained results and the comparison with the traditional association rule metrics have shown (i) the capability of the proposed metric of allowing the identification of actual FDEPs, (ii) the ability to discriminate spurious rules, and (iii) the robustness with respect to statistical fluctuations.

For future works, one direction lies in the application of the proposed metric to the direct identification of FDEPs without the need of generating association rules. Also, since the concept of dependency plays an important role in many fields, such as database design, machine learning, knowledge discovery and medical applications, the application of the proposed metric for other objectives than the identification of FDEPs in CTIs will be investigated.

Funding Open access funding provided by Politecnico di Milano within the CRUI-CARE Agreement.

Declarations

Conflict of interest The authors declare that they have no conflict of interest.

References

Agrawal R, Imieliński T (1993) Mining association rules between sets of items in large databases. ACM SIGMOD Rec 22(2):207–216

Antonello F, Baraldi P, Shokry A, Zio E, Gentile U, Serio L (2020) Data-driven extraction of association rules of dependent abnormal behaviour groups. In: Proceedings of the 29th European safety and reliability conference, ESREL 2019, 2020, pp 3308–3313

Antonello F, Baraldi P, Gentile U, Serio L, Shokry A, Zio E (2020) A method for inferring causal dependencies among abnormal behaviours of components in complex technical infrastructures. In: Proceedings of the 30th European safety and reliability conference and the 15th probabilistic safety assessment and management conference, 2020, pp 309–316

Antonello F, Baraldi P, Shokry A, Zio E, Gentile U, Serio L (2021a) Association rules extraction for the identification of functional dependencies in complex technical infrastructures. Reliab Eng Syst Saf 2021(209):107305

Antonello F, Baraldi P, Serio L, Zio E (2021b) A novel association rule mining method for the identification of rare functional dependencies in complex technical infrastructures from alarm data. Exp Syst Appl 2021(170):114560

Antonello F, Baraldi P, Zio E, Serio L (2022) A niching augmented evolutionary algorithm for the identification of functional dependencies in complex technical infrastructures from alarm data. IEEE Syst J. https://doi.org/10.1109/JSYST.2022.3146014

Ballantyne A, Lawrance N, Small M, Hodkiewicz M, Burton D (2018) Fault prediction and modelling in transport networks. 2018 IEEE international symposium on circuits and systems (ISCAS), Florence, pp 1–5

Benites F, Sapozhnikova E (2014) Evaluation of hierarchical interestingness measures for mining pairwise generalized association rules. IEEE Trans Knowl Data Eng 26(12):3012–3025

Billinton R, Allan RN (1992) Network modelling and evaluation of complex systems. Reliability evaluation of engineering systems. Springer, Boston

Cantelmi R, Di Gravio G, Patriarca R (2021) Reviewing qualitative research approaches in the context of critical infrastructure resilience. Environ Syst Decis 41:341–376

Chang SE, McDaniels TL, Mikawoz J, Peterson K (2007) Infrastructure failure interdependencies in extreme events: power outage consequences in the 1998 Ice storm. Nat Hazards 41(2):337–358

Del Jesus MJ, Gámez JA, González P, Puerta JM (2011) On the discovery of association rules by means of evolutionary algorithms. Wiley Interdiscipl Rev 1(5):397–415

Etesami J, Kiyavash N (2017) Measuring causal relationships in dynamical systems through recovery of functional dependencies. IEEE Trans Signal Inf Process Netw 3(4):650–659

Ghosh A, Nath B (2004) Multi-objective rule mining using genetic algorithms. Inf Sci (NY) 163(1–3):123–133

Geng L, Hamilton HJ (2006) Interestingness measures for data mining: a survey. ACM Comput Surv 38:2006

Hämäläinen W, Webb GI (2019) A tutorial on statistically sound pattern discovery. Data Min Knowl Disc 33:325–377

Hui CY et al (2005) Mining quantitative associations in large database. Softw Eng Middlew 3399(60373053):405–416

Eusgeld I, Nan C, Dietz S (2011) System-of-systems approach for interdependent critical infrastructures. Reliab Eng Syst Saf 96(6):679–686

Luna JM, Ondra M, Fardoun HM, Ventura S (2018) Optimization of quality measures in association rule mining: an empirical study. Int J Comput Intell Syst 12(1):59–78

Kröger W, Zio E (2011) Vulnerable systems. Springer, London

Hickford AJ, Blainey SP, Ortega Hortelano A et al (2018) Resilience engineering: theory and practice in interdependent infrastructure systems. Environ Syst Decis 38:278–291

Li Y, Liu J (2018) A Bayesian network approach for imbalanced fault detection in high speed rail systems. In: 2018 IEEE international conference on prognostics and health management (ICPHM), Seattle, WA, pp 1–7

Jin C, Rong L, Sun K (2017) Modeling of interdependent critical infrastructures network in consideration of the hierarchy. Knowledge and systems sciences. Springer, Singapore, pp 117–128

Ji Y, Ying H, Tran J, Dews P, Mansour A, Michael Massanari R (2013) A method for mining infrequent causal associations and its application in finding adverse drug reaction signal pairs. IEEE Trans Knowl Data Eng 25(4):721–733

Johansson J, Hassel H (2010) An approach for modelling interdependent infrastructures in the context of vulnerability analysis". Reliab Eng Syst Saf 95(12):1335–1344

Marinica C, Guillet F (2010) Knowledge-based interactive postmining of association rules using ontologies. IEEE Trans Knowl Data Eng 22(6):784–797

Mathu T, Narmadha D, Geetha S (2011) Mining and post-mining of time stamped association rules. In: 3rd international conference on Electronics Computer Technology (ICECT) 2011, vol 4, pp 149–153

Mendonça D, Wallace WA (2006) Impacts of the 2001 world trade center attack on New York city critical infrastructures. J Infrastruct Syst 12(4):260–270

Mosleh A (1991) Common cause failures: an analysis methodology and examples. Reliab Eng Syst Saf 1991(34):249–292

Moura MDC, Lins IS, Droguett EL, Soares R, Pascual R (2015) A multi-objective genetic algorithm for determining efficient risk-based inspection programs. Reliab Eng Syst Saf 133(2015):253–265

Mukhopadhyay A, Maulik U, Bandyopadhyay S, Coello CAC (2014) Survey of multiobjective evolutionary algorithms for data mining: Part II. IEEE Trans Evol Comput 18(1):25–35

Nielsen, Serio L (2016) Technical services: unavailability root causes, strategy and limitations. In: Proceedings on 7th Evian Workshop on LHC beam operation, Evian Les Bains, France, December 2016

O'Connor A, Mosleh A (2016) A general cause based methodology for analysis of common cause and dependent failures in system risk and reliability assessments. Reliab Eng Syst Saf 145:341–350

Piatetsky-Shapiro G (1991) Discovery, analysis and presentation of strong rules. In: Piatetsky-Shapiro G, Frawley W, eds, Knowledge discovery in databases, pp 229–248. AAAI Press

Sánchez D, Serrano JM, Blanco I et al (2008) Using association rules to mine for strong approximate dependencies. Data Min Knowl Disc 16:313–348

Serio L, Antonello F, Baraldi P, Castellano A, Gentile U, Zio E (2018) Smart framework for the availability and reliability assessment and management of accelerators technical facilities. In: 9th International Particle Accelerator Conference, IPAC 2018

Srikant R, Agrawal R (1996) Mining quantitative association rules in large relational tables. ACM SIGMOD Rec 25(2):1–12

Su H, Zio E, Zhang J, Li X (2018) A systematic framework of vulnerability analysis of a natural gas pipeline network. Reliab Eng Syst Saf 175:79–91

Todd et al (2016) LHC Availability 2016: Standard Proton Physics. CERN, Geneva, Switzerland, Rep. CERN-ACC.NOTE-2016–0067, December 2016

Wen RZ, Sun BT, Zhou BF (2010) Field survey of Mw8, 8 Feb 27. Chile earthquake and tsunami. Adv Mater Res 250(2011):2102–2106

Witten IH, Frank E (2016) Data mining: practical machine learning tool and techniques. Morgan, New York

U.S.-Canada Power System Outage Task Force (2004) Final report on The August 14, 2003 Blackout in The United States and Canada: Causes and Recommendations

Van Leeuwen M, Galbrun E (2015) Association discovery in two-view data. IEEE Trans Knowl Data Eng 27(12):3190–3202

Yao H, Hamilton HJ (2008) Mining functional dependencies from data. Data Min Knowl Disc 16:197–219

Zaki MJ (2000) Generating non-redundant association rules. Proceedings of the sixth ACM SIGKDD international conference on Knowledge discovery and data mining - KDD '00: 34–43

Zhang A, Shi W, Webb GI (2016) Mining significant association rules from uncertain data. Data Min Knowl Disc 30:928–963

Zio E (2009) Computational methods for reliability and risk analysis. World Scientific, Singapore

Zio E (2016a) Some challenges and opportunities in reliability engineering. IEEE Trans Reliab 99:1769–1782

Zio E (2016b) Challenges in the vulnerability and risk analysis of critical infrastructures. Reliab Eng Syst Saf 152(2016):137–150