Revista Brasileira de Computação Aplicada, April, 2019

DOI: 10.5335/rbca.v11i1.8820
Vol. 11, Nº 1, pp. 59–66
Homepage: seer.upf.br/index.php/rbca/index

ISSN 2176-6649

ORIGINAL PAPER

Probabilistic logic reasoning for subjective interestingness analysis

José Carlos F. da Rocha[1], Alaine M. Guimarães[1] and Valter L. Estevam Jr.[2]

[1]UEPG and [2]IFPR

*jrocha@uepg.br; amguimaraes@uepg.br; †valter.junior@ifpr.edu.br.

Received: 2018-11-02. **Revised:** 2019-02-08. **Accepted:** 2019-03-13.

Abstract

This paper presents an approach that uses probabilistic logic reasoning to compute subjective interestingness scores for classification rules. In the proposed approach, domain knowledge is represented as a probabilistic logic program that encodes information from experts and statistical reports. The computation of interestingness scores is performed by a procedure that applies linear programming to reasoning regarding the probabilities of interest. It provides a mechanism to calculate probability-based subjective interestingness scores. Further, a sample application illustrates the use of the described approach.

Key words: Interestingness analysis; KDD; Probabilistic inference.

Resumo

Este trabalho apresenta uma abordagem que utiliza a inferência em lógica probabilística para calcular escores de interessabilidade subjetiva de regras de classificação. Na abordagem proposta, o conhecimento do domínio é representado como um programa em lógica probabilística que contém informações fornecidas por especialistas ou extraídas de relatórios estatísticos. O cômputo dos escores de interessabilidade é executado por um procedimento que emprega a programação linear para inferir o valor de probabilidades de interesse. Isto fornece um mecanismo para calcular escores probabilísticos para a interessabiliadade subjetiva. Um exemplo de aplicação ilustra a utilização da abordagem descrita.

Palavras-Chave: Análise de interessabilidade; KDD; Inferência probabilística.

1 Introduction

Knowledge discovery in databases (KDD) is a field of computer science that investigates the theoretical basis of transforming raw data into useful and comprehensive information, and develops computational methods to achieve the same. The aim is to identify patterns that encode information that could be useful for solving a target problem. Such a process is usually abstracted into a three-step procedure: data preprocessing, data mining, and evaluation and interpretation of discovered patterns.

A knowledge discovery task that is often addressed with the use of KDD techniques is the development of classifier systems. Here, KDD algorithms inspect a data set to determine patterns that facilitate the building of a function that relates the category of an object to its characteristics/attributes. In this context, the classification rule mining aims at discovering implication patterns where the antecedent represents a logical constraint on the values of attributes used to describe an object, and the consequent specifies a label that identifies the class of the object. If the description of an object matches the condition in the left side of a rule, it is classified into the respective category.

The success of rule mining is primarily evaluated by identifying a set of rules that allow the implementation of an accurate classifier. Additionally, discovered rules may also be subject

to an interestingness analysis. A KDD step determines if the mined patterns are worthwhile, i.e., if they represent any novel, useful, valid, and understandable knowledge (Han; 2005). The output of an interestingness analysis process thus associates each discovered pattern with a number of scores, which measure the relevance of a rule for the application targets.

In particular, subjective interestingness analysis estimates the relevance of a pattern given the domain knowledge, user beliefs, and task goals. It generally requires the implementation of a knowledge base that stores information, allowing evaluation if mined patterns are pertinent. Domain knowledge is often uncertain, and hence, it is usually necessary to employ a reasoning scheme to address uncertainty during the interestingness analysis. Considering that, this work presents a probabilistic approach for subjective analysis. The proposed approach defines a scheme that allows one to represent uncertain knowledge about some application domain propositions and provides a procedure to execute the inferences and calculate probability-based interestingness measures.

The rationale underlying the proposed approach is to use probabilistic logic to encode the domain knowledge into a knowledge base (KB), and an associated reasoning procedure to compute interestingness measures. The assertions in the knowledge base are assumed to represent knowledge elicited from expert beliefs, inferred from descriptive statistics, or obtained from fitted models and correlation data. Imprecise probabilistic assignments are dealt with as interval-valued probabilities. The reasoning procedure makes use of linear programming.

A sample application illustrates the use of the proposed procedure for computing two interestingness scores—self-information and level—for a set of rules generated by the JRIP algorithm (Cohen; 1995) on the UCI Breast Cancer Data Set. The self-information evaluates whether a rule is unexpected (Bie; 2011), while the level of interest is a robust measure of predictive accuracy (Gay and Boullé; 2013).

This article is organized as follows: Section 2 presents the background on probabilistic logic, classification rule mining, interestingness analysis and interestingness measures. Section 3 presents the proposed approach. Section 4 illustrates the use of the proposed method through an application example. Section 5 discusses the main issues related to the use of the proposed approach in interestingness analysis. The last section presents the final remarks of this study.

2 Background review

Classification rule mining aims at discovering a set of implication patterns that relates certain object features (attributes) to a label representing the category of the object under analysis (Vashishtha et al.; 2011). Let $\mathbf{X} = \{X_1 \ldots, X_n\}$ denote a set of variables whose elements identify the attributes[1] used to describe the objects to be classified. The sample space of X_i is denoted by Ω_i. Furthermore, let C be a categorical variable whose sample space, Ω_C, enumerates every classification hypothesis. Given a data set $\mathbf{D}$ with m instances of the form $(X_1 \ldots, X_n, C)$, a classification rule mining algorithm applies inductive learning methods to identify a collection of logical expressions of the form $F_1 \wedge F_2 \cdots \wedge F_t \rightarrow H$ (Fürnkranz et al.; 2014). Each F_i symbolizes an expression defined on the elements of $\mathbf{X}$, and H stands for a class assignment $C = c$ such that $c \in \Omega_C$. This work assumes that each F_i is an expression $X_j \odot x_{j,k}$, where $x_{j,k} \in \Omega_j$, $\odot$ is an relational operator from the set $\{<, >, \leq, \geq, =\}$, and $1 \leq t \leq n$.

Classification rule mining performance is primarily assessed by the accuracy of the classifier constructed on the discovered rules. Additionally, sometimes it may be convenient to evaluate if the mined patterns are also valid, novel, useful, and understandable (Geng and Hamilton; 2006; McGarry; 2005). This type of investigation is called interestingness analysis, and it aims at computing measures that quantify how interesting a pattern is from an objective or subjective point of view. In objective interestingness analysis, pattern evaluation is based on statistical measures that estimate the strength that the data provides to the pattern. Subjective analysis, on the other hand, intends to appraise if the discovered rule meets user beliefs and objectives, as well as fits to data (Leeuwen et al.; 2016). Generally, interestingness analysis is a post-processing step, and hence, the scores are used to filter or rank the rules.

This work considers two subjective interestingness measures: the self-information and the interestingness level. Let R be a rule $F_1 \wedge F_2 \cdots \wedge F_t \rightarrow H$. The self-information of R is defined as (Bie; 2011):

$$I(R) = -\log_2(P(R)) \tag{1}$$

Self-information, also known as surprisal, quantifies how expected a pattern is. A value approximately equal to zero indicates that R appears highly plausible considering a probability distribution p, defined over the sentences in a knowledge base. On the other hand, the higher $I(R)$ is, the more surprising (unexpected or improbable) R is.

The level measure, denoted by $level(R)$ is a Bayesian score that weighs the posterior probability of R by the posterior probability of a default rule R_0[2] (Egho et al.; 2015; Gay and Boullé; 2013). $level(R)$ is expressed as follows:

$$level(R) = 1 - \frac{c(R)}{c(R_0)}. \tag{2}$$

In Expression (2), $c(R)$ is the cost of R. It is the negative logarithm of $p(\mathbf{D} \wedge R)$ which yields $c(R) = -\log_2 p(\mathbf{D} \wedge R) = -\log_2 P(\mathbf{D}|R) - \log_2 P(R))$. As $P(R|\mathbf{D}) \propto p(\mathbf{D} \wedge R)$, $c(R)$ is related to the posterior probability of R given the data. $c(R_0)$ is the cost of a

[1] In this work, it is assumed that a variable can be categorical, discrete, or continuous.

[2] A default rule R_0 has no antecedents (its form is $\rightarrow H$).

default rule, i.e., $c(R_0) = -\log_2(P(\mathbf{D}|R_0)) - \log_2(P(R_0))$.

The logic underlying the level score is that, by exploring the posterior probabilities, this measure provides information that enables simultaneous evaluation of the data fitting and prior expectancy. In addition, it is a normalized score (upper bounded in 1), which allows comparison of the performance of R with that of the default rule. Fundamentally, if:

- $level(R) \leq 0$, the rule is not interesting because it has equal or less probability than R_0;
- $level(R) = 1$, the rule exactly fits the observations and prior beliefs;
- $0 < level(R) < 1$, it indicates rules with a certain degree of interestingness.

The $level(\cdot)$ measure assigns a higher score to a rule if it is more likely than the default rule.

2.1 Propositional probabilistic logic

Propositional logic represents categorical facts by means of formulas defined on propositional variables (Russell and Norvig; 2010). Let *true* and *false* be two constant values, and let $\mathbf{V} = \{v_1, ...v_m\}$ be a set of propositional variables. The elements of $\mathbf{V}$ are named atomic formulas and can assume one of the two constant values. The compound formulas are denoted by $S_1, S_2, and \dots S_m$, and are constructed by connecting an atomic or a compound formula to another by means of the logical operators $\wedge, \vee, \neg$, and $\rightarrow$. A compound formula is also *true* or *false*, and its value is a function of the truth assignment to its variables and the semantics of the operators. A *truth assignment*, w, is a vector that assigns either *true* or *false* to each propositional variable in a formula. This work assumes the usual semantics for operators (Hamilton; 1988).

Probabilistic logic extends propositional logic in order to allow the treatment of uncertain knowledge (Hansen and Perron; 2008). Probabilistic logic thus assigns a probability measure π_i to every formula S_i such that the statement $P(S_i) = \pi_i$ expresses the belief of an agent on S_i. If some agent's beliefs are imprecise, they can be expressed by inequalities such as $P(S_i) \geq \pi_i$ or $P(S_i) \leq \pi_i$ or by interval probability statements such as $\underline{\pi}_i \leq P(S_i) \leq \overline{\pi}_i$. Here, $\underline{\pi}_i$ and $\overline{\pi}_i$ are the lower and upper bounds of π_i, respectively. The conditional statements expressing the belief on S_i given an event S_j can be written as $P\left(S_i|S_j\right) = \pi_{i,j}$, $P\left(S_i|S_j\right) \geq \pi_{i,j}$, $P\left(S_i|S_j\right) \leq= \pi_{i,j}$, or $\underline{\pi}_{i,j} \leq P\left(S_i|S_j\right) \leq \overline{\pi}_{i,j}$.

A probabilistic logic knowledge base $\mathbf{K}$ is a collection of probabilistic logic sentences on $\mathbf{V}$. It can be considered as a pair $(\mathbf{S}, \mathbf{\Pi})$, where $\mathbf{S}$ is a set of propositional sentences $\{S_1 \dots, S_m\}$ associated with the probability assignments $\mathbf{\Pi}$. $\mathbf{\Pi}$ can be partitioned as $\begin{pmatrix} \mathbf{\Pi_1} \\ \mathbf{\Pi_2} \\ \mathbf{\Pi_3} \end{pmatrix}$; $\mathbf{\Pi_1}$, $\mathbf{\Pi_2}$, and $\mathbf{\Pi_3}$ are column vectors specifying the equality constraints ($P(S_i) = \pi_i$), lower bounds ($P(S_i) \geq \pi_i$), and upper bounds ($P(S_i) \leq \pi_i$), respectively.

Let $\mathbf{M}$ be the set of all possible truth assignments w_j of $\mathbf{V}$ and p_j be the probability of w_j in the joint distribution associated with $\mathbf{V}$. It holds that (Hooker; 1992):

$$P(S_i) = \sum_{w:w_j \in \mathbf{M} \wedge m(S_i,w_j)} p_j = \mathbf{a}_i^{\intercal}\mathbf{p}. \quad (3)$$

Here, $m(S_i, w_j)$ indicates that w_j is a model for S_i, $\mathbf{p} = (p_1 \dots, p_{2^n})^{\intercal}$ is a vector on the joint probability of $\mathbf{V}$, and $\mathbf{a}_i$ denotes a vector whose j^{th} element is 1 if S_i is true in w_j and zero otherwise. A knowledge base $\mathbf{K}$ is said to be consistent if its sentences are consistent with the axioms of probability theory.

Equation 3 can be used to build a procedure for solving the inferences (Hooker; 1992) about the lower and upper probabilities of a target sentence. Let S be the target, $\mathbf{K}$ be a knowledge base, and $\underline{P}(S)$ and $\overline{P}(S)$ be the lower and upper probabilities of S, respectively. $\underline{P}(S)$ and $\overline{P}(S)$ can be defined as a linear function $\mathbf{a}^{\intercal}\mathbf{p}$, which must be minimized or maximized for a given number of constraints derived from $\mathbf{K}$. This work assumes that the knowledge base constraints can be grouped into three matrices, namely $A_{m_1 \times 2^N}$, $A_{m_2 \times 2^N}$, and $A_{m_3 \times 2^N}$. These matrices store the linear expressions related to the equality, less than or equal, and greater than or equal constraints, respectively. It generates the following linear program:

$$\begin{array}{lll} \min / \max & \mathbf{a}^T\mathbf{p} & \\ s.t & & \\ & A_{l_1 \times 2^N} & = \mathbf{\Pi_1} \\ & A_{l_2 \times 2^N} \quad \times \mathbf{p} & \geq \mathbf{\Pi_2} \\ & A_{l_3 \times 2^N} & \leq \mathbf{\Pi_3}. \end{array}$$

Furthermore, $\mathbf{p} \geq \mathbf{0}$ and $\mathbf{1}^{\intercal}\mathbf{p} = 1$.

It must be noted that conditional statements can also be represented in probabilistic logic. For example, let $P(S_1|S_2)$ be the probability of a statement S_1 conditional to S_2. The expressions $P(S_1|S_2) = \pi_{1,2}$, $P(S_1|S_2) \geq= \pi_{1,2}$, and $P(S_1|S_2) \leq= \pi_{1,2}$ express constraints on that belief. These expressions yield a number of linear equations/inequalities as follows: $P(S_1 \wedge S_2) - P(S_2) \cdot \pi_{1,2} = 0$, $P(S_1 \wedge S_2) - P(S_2) \cdot \pi_{1,2} \geq 0$ $P(S_1 \wedge S_2) - P(S_2) \cdot \pi_{1,2} \leq 0$, respectively.

3 A probabilistic logic approach for interestingness analysis

This work assumes that the data mining team intends to construct a knowledge-base, $\mathbf{K}$, to analyze the interestingness of certain classification rules. This paper embraces such an approach by exploring probabilistic logic to represent domain knowledge and related reasoning procedures to support the computation of probability-based interestingness measures. More specifically, let R be the sentence that symbolizes a classification rule to be analyzed such that $R \equiv F_1 \wedge F_2 \cdots \wedge F_t \rightarrow H$. Let it also be that $P(S_1) \dots P(S_m)$ are the sentences in $\mathbf{K}$. The propositional components of those sentences denote facts and associations relative to the terms that appear in the rule.

After developing such a knowledge base, it is possible to proceed as in section 2.1 and to state

a problem of probabilistic logic inference whose assignments represent the uncertainty about facts and relationships that the team believes to be relevant to the calculation of $P(R)$. Equation 4 illustrates the structure of the inference problem whose solution gives the lower and upper bounds of $P(R)$.

$$\begin{array}{ll} \min / \max & P(R) \\ s.t. & P(S_1) = \pi_1 \\ & \dots \\ & P(S_m) = \pi_t \\ & P(H) \end{array} \tag{4}$$

Example 1 is a simple straightforward application of the proposed approach.

Example 1: Let X_1 and X_2 be two normally distributed variables such that $X_1 \sim N(1; 0.1)$ and $X_2 \sim N(4; 1)$ and $H \equiv (C = c_1)$. Assume that the prevalence of class c_1 is higher than or equal to 0.8. Given a rule $R \equiv (X_1 \leq 0.901 \wedge X_2 \leq 5 \rightarrow C = c_1)$, it is possible to use the previous information to construct a probabilistic logic program for computing $\underline{P}(R)$ and $\overline{P}(R)$. Let $P(S_1) = P(X_1 \leq 0.901) = 0.16$, $P(S_2) = P(X_2 \leq 5 \rightarrow C = c_1) = 0,84$. Additionally, let $P(S_0) = 0.6$ be the marginal probability of H and $P(S_2|S_0) = 0.7$. The upper and lower bounds for $P(R)$ are obtained by solving the following linear program:

$$\begin{array}{ll} \min / \max & \mathbf{a}^\intercal \mathbf{p} \\ s.t & \mathcal{A} \times \mathbf{p} = \Pi \\ & \mathbf{1}^\intercal \mathbf{p} \\ & p_i \geq 0, i = 1..8 \end{array} \tag{5}$$

where, $\mathcal{A} = \begin{pmatrix} \mathbf{a}_0 & \mathbf{a}_1 & \mathbf{a}_2 & \mathbf{a}_{2|0} \end{pmatrix}^T$, $\mathbf{p} = \begin{pmatrix} p_1 & \dots & p_8 \end{pmatrix}^T$, and $\Pi = \begin{pmatrix} 0.65 \\ 0.16 \\ 0.84 \\ 0 \end{pmatrix}$. The rows in $\mathcal{A}$ are defined as $\mathbf{a}_0 = (1, 0, 1, 0, 1, 0, 1, 0)$, $\mathbf{a}_1 = (1, 1, 1, 1, 0, 0, 0, 0)$, $\mathbf{a}_2 = (1, 1, 0, 0, 1, 1, 0, 0)$ and $\mathbf{a}_{2|0} = (0.3, 0, -0.7, 0, 0.3, 0, -0.7, 0)$. The objective function is $\mathbf{a} = (1, 0, 1, 1, 1, 1, 1, 1)$.

The example above makes it evident that using the proposed approach relies on a knowledge engineering step that aims to acquire knowledge regarding logical associations among domain variables and to elicit their respective probabilities. However, knowledge acquisition is a hard task (Russell and Norvig; 2010). So, for the sake of simplicity, at first, it is supposed that the density/distribution $p\left(X_j\right)$ is known for each $X_j \in \mathbf{X}$.

Of course, the analysis team may have to consult several sources of information to get the probability of each sentence. The probabilities can be encoded into statistical reports (Barbaros et al.; 2014; van der Gaag et al.; 2013; Sivia and Skilling; 2006). Its elicitation may demand meta-analysis of scientific and technical literature (Garthwaite et al.; 2005) or knowledge acquisition from experts (O'Hagan et al.; 2006). All of these strategies couple with the condition stated above and the linear program in Equation 5.

If the available data and experts do not support an exact probabilistic assignment for every sentence, the analysis team could extend the model by using a formalism based on the imprecise probability theory (Levi; 1980; Walley; 1991). Imprecise probabilities also make possible to deal with situations where the available information is in the form of comparative probability statements. For example, let Q_1, Q_2, and Q_3 be three composed sentences defined on $S_1 \dots, S_t$ such that experts are aware that: (a) Q_1 is as or more probable than Q_2, and (b) Q_3 is as probable or more probable than Q_1. Thus, $P(Q_1) \geq P(Q_2)$ and $P(Q_3) \geq P(Q_1)$ can be added to the program.

Similarly, if it is known that $k_1P(Q_1) \leq k_2P(Q_2)$, $k_1P(Q_1) \geq k_2P(Q_2)$, or $P(Q_1) = P(Q_2)$ for $k_1, k_2 \in \mathbb{R}$, the expressions $k_1P(Q_1) - k_2P(Q_2) \leq 0$, $k_1P(Q_1) - k_2P(Q_2) \geq 0$, or $P(Q_1) - P(Q_2) = 0$, respectively, could be added into the probabilistic program. As before, such constraints can be rewritten using a vectorial notation as follows: $\mathbf{b}_{1,2} = k_1\mathbf{b}_1 - k_2\mathbf{b}_2 \odot 0$, such that $\mathbf{b}_1$ and $\mathbf{b}_2$ are the row vectors relative to $P(Q_1)$ and $P(Q_2)$. Qualitative constraints can be grouped into a system $\mathcal{B} \times \mathbf{p} \odot \Pi$ and further can be appended to program 5 as follows:

$$\begin{array}{ll} \min / \max & \mathbf{c}^\intercal \mathbf{p} \\ s.t & \begin{bmatrix} \mathcal{A} \\ \mathcal{B} \end{bmatrix} \times \mathbf{p} \odot \begin{bmatrix} \Pi \\ \mathbf{0} \end{bmatrix} \\ & \mathbf{1}^\intercal \mathbf{p} = \mathbf{1} \\ & p_i \geq 0, i = 1..2^t. \end{array} \tag{6}$$

3.1 Integrating information about correlation

There can exist a case wherein the correlation data between X_i and X_j is handy. It could be useful to explore that information in order to constrain the probabilistic relationship between those variables. Berleant and Jianzhong (2004) and Berleant et al. (2007) present a procedure to calculate envelopes for the joint probability distribution of two variables, X_i and X_j, given their Pearson correlation. This section describes the use of that procedure in order to discover the lower and upper limits, $\underline{\pi}_{i\wedge j}$ and $\overline{\pi}_{i\wedge j}$, for a sentence $P\left(S_i \wedge S_j\right)$; here, S_i and S_j indicate that the values of X_i and X_j belong to a given interval in Ω_i and Ω_j.

Let X_i and X_j be two continuous attributes, $p(X_i)$ and $p\left(X_j\right)$ their densities, and r the correlation between them. Further, let Z and Y be two variables whose values $z_1 \dots z_{n_1}$ and $y_1 \dots y_{n_2}$, respectively, are obtained with the discretization of X_i and X_j into n_1 and n_2 bins. The sample spaces of Z and Y are denoted by Ω_z and Ω_y. In addition, let $p(Z)$ and $p(Y)$ be the marginal distributions of those new variables. The entries $P(z_k)$ and $P(y_l)$ can be calculated from $p(X_i)$ and $p\left(X_j\right)$ by doing

$$P(z_k) = P\left(\underline{x}_{i,k} < X_i \leq \overline{x}_{i,k}\right)$$

and

$$P(y_l) = P\left(\underline{x}_{j,l} < X_j \leq \overline{x}_{j,l}\right).$$

Here $\underline{x}_{i,k}$ and $\overline{x}_{i,k}$ ($\underline{x}_{j,l}$ and $\overline{x}_{j,l}$) are the limits of the k^{th} (l^{th}) bin of Z (Y).

Let there be a case where S_i and S_j appear in the

antecedent of a classification rule such that $S_i \equiv (X_i \geq x_{i,a})$ and $S_j \equiv (X_j \geq x_{j,b})$. Without loss of generality, assume that $x_{i,a}$ and $x_{j,b}$ are the lower bounds of the intervals which define the bins z_a and y_b in Ω_Z and Ω_Y. Hence, $P(S_i)$ and $P\left(S_j\right)$ can be expressed in terms of $p(Z)$ and $p(Y)$. Marginalization of $p(Z,Y)$ allows to define the next equations:

$$\begin{aligned} P(S_i) &= \sum_{k=a}^{n_1} \sum_{l=1}^{n_2} P(Z = z_k \wedge Y = y_l) = \pi_i \\ P\left(S_j\right) &= \sum_{l=b}^{n_2} \sum_{k=1}^{n_1} P(Z = z_k \wedge Y = y_l) = \pi_j \end{aligned} \tag{7}$$

Similarly, $P\left(S_i \wedge S_j\right)$ can be formulated in terms of $p(Z,Y)$ as follows:

$$P\left(S_i \wedge S_j\right) = \sum_{t_* \in \mathbf{t}} P(Z = z_{t_*} \wedge Y = y_{t_*}) = \pi_{i \wedge j}. \tag{8}$$

In this expression, $\pi_{i\wedge j}$ denotes the unknown value $P\left(S_i \wedge S_j\right)$, and $\mathbf{t}$ is a vector of pairs of indexes such that for all $t_* = (k,l) \in \mathbf{t}$, the intervals represented by z_k and y_l are consistent with the sentence $S_i \wedge S_j$. Equations 7 and 8 relate the joint $p\left(X_i, X_j\right)$ to the sentences $P\left(S_i \wedge S_j\right)$, $P(S_i)$, and $P\left(S_j\right)$ through $P(Z,Y)$. As before, those equations can be represented in a vector form and appended to the Program (6).

However, the usability of those constraints depends on an estimate or bounds for $\pi_{i\wedge j}$. Following Berleant et al. (2007) and Berleant and Jianzhong (2004) that bounds can be computed from r and $p(X_i)$ and $p\left(X_j\right)$, with equations 9 and 10 :

$$\sum_{k,l}^{n_1,n_2} \underline{z_k y_l} P(Z = z_k \wedge Y = y_l) \geq \underline{\mu_i \mu_j} + r\sqrt{\underline{\sigma_i^2 \sigma_j^2}} \tag{9}$$

$$\sum_{k,l}^{n_1,n_2} \underline{z_k y_l} P(Z = z_k \wedge Y = y_l) \leq \overline{\mu_i \mu_j} + r\sqrt{\overline{\sigma_i^2 \sigma_j^2}} \tag{10}$$

The equations presented by Berleant and Jianzhong (2004) allow to obtain an outer envelope for $p(Z,Y)$ given the correlation data along with the upper and lower bounds for the mean and variance of X and Y. Moreover, equations 7, 8, 9, and 10 can be grouped in the form of a linear system $\mathcal{D}$. If appended to program (6), $\mathcal{D}$ defines additional constraints in the optimization program and hence, can contribute to obtaining tighter intervals for $P(R)$.

In particular, the utilization of correlation data demands the acquisition of $\underline{\mu_j}$, $\overline{\mu_j}$, $\underline{\sigma_i^2}$, $\underline{\sigma_j^2}$, $\overline{\sigma_i^2}$, and $\overline{\sigma_j^2}$. As proposed by Berleant and Jianzhong (2004) and Berleant et al. (2007), this work assumes that these limits are entered by the analysis team or calculated by interval optimization upon $P(Z)$ and $P(Y)$.

3.2 Evaluating interestingness

The described approach assumes that interestingness analysis is performed after the data mining step (i.e., a post-processing step) and aims at sorting the rules by surprisal or level of interest scores. The self-information of a rule R can be obtained by solving the linear programs described in section 3. They produces lower and upper probability estimates for $P(R)$ and a respective interval $[\underline{I}(R), \overline{I}(R)]$ for the self-information of R. Here

$$\underline{I}(R) = -\log_2(\overline{P}(R)) \qquad \overline{I}(R) = -\log_2(\underline{P}(R)) \tag{11}$$

If $P(R)$ is an interval-valued probability, $c(R)$, the numerator of $level(R)$ is also an interval $[\underline{c}(R), \overline{c}(R)]$ whose extremes are:

$$\underline{c}(R) = -log(\overline{P}(R)) - log(P(\mathbf{D}|R)) \tag{12}$$

$$\overline{c}(R) = -log(\underline{P}(R)) - log(P(\mathbf{D}|R)) \tag{13}$$

Again, it induces an interval on the level score whose limits $\underline{level}(R)$ and $\overline{level}(R)$] are

$$\underline{level}(R) = 1 - \frac{\overline{c}(S)}{c(S_0)} \qquad \overline{level}(S) = 1 - \frac{\underline{c}(S)}{c(S_0)} \tag{14}$$

After obtaining the interval for $level(R)$, the interestingness analysis continues by inspecting its lower and upper bounds. If $\underline{level}(R)$ is greater than 0, it implies that the rule appears to be interesting given prior knowledge as well as effective in describing data even if it was computed on the lower bound for $P(S)$. On the other hand, $\overline{level}(R) < 0$ indicates that in the light of the background information, the rule is not interesting even if the analysis considers an upper bound for $P(S)$. Finally, if $0 \in [\underline{level}(S), \overline{level}(S)]$, no conclusion can be drawn about the robustness of the rule.

4 An example

In this section, the proposed approach is used to carry out interestingness analysis on three classification rules learned by the JRIP algorithm (Cohen; 1995) from the Breast Cancer Wisconsin Data Set (Wolberg. et al.; 1994). Before the learning step, the data set was split into two partitions: a training data set with 379 cases and a test data set with 190 cases. JRIP generated the next three rules:

- rule (a): (concave points n1 $\geq$ 0.05182) and (perimeter n3 $\geq$ 113.9) $\rightarrow$ Diagnosis=malign;
- rule (b): (concave points n1 $\geq$ 0.05839) and (texture n3 $\geq$ 23.75) $\rightarrow$ Diagnosis=malign;
- rule (c): (radius n3 $\geq$ 15.65) and (texture n3 $\geq$ 28.06) and (smoothness n3 $\geq$ 0.1094) $\rightarrow$ Diagnosis=malign.

The left side of those rules refers to certain features of a cellular nucleus and the right side shows a class label. The propositions that constitute the rules were denoted as: $S_0 \equiv$ (Diagnosis=malign), $S_1 \equiv$ (concave points n1 $\geq$ 0.05182), $S_2 \equiv$ (perimeter n3 $\geq$ 113.9), $S_3 \equiv$ (radius n3 $\geq$ 15.65), $S_4 \equiv$ (texture n3 $\geq$ 28.06),

$S_5 \equiv$ (radius n3 $\geq$ 15.65), $S_6 \equiv$ (texture n3 $\geq$ 28.06), and $S_7 \equiv$ (smoothness n3 $\geq$ 0.1094). The rules (a), (b), and (c) were associated to the sentences R_a, R_b, and R_c so that $R_a \equiv S_1 \wedge S_2 \rightarrow S_0$, $R_b \equiv S_3 \wedge S_4 \rightarrow S_0$, and $R_c \equiv S_5 \wedge S_6 \wedge S_7 \rightarrow S_0$.

The following sentences was entered into the knowledge base: $P(S_1) \geq 0.4$, $P(S_2) \geq 0.33$, $P(S_3) = 0.34$, $P(S_4) = 0.58$, $P(S_5) = 0.499$, $P(S_6) = 0.34$, and $P(S_7) = 0.845$. $P(R_0)$ was set to 0.0008, the prevalence of breast cancer in the US[3]. The marginal sentences were followed by the next conditional probability sentences: $P(S_1|S_0) \geq 0.51$, $P(S_2|S_0) \geq 0.82$, $P(S_4|S_0) = 0.85$, and $P(S_6|S_0) = 0.44$. The sentences in KB were elicited from an hypothetical expert[4].

The knowledge base also had a number of qualitative constraints on the joint probabilities of S_1 and S_2: $P(S_1 \wedge S_2) \geq P(S_1 \wedge \neg S_2)$, $P(S_1 \wedge S_2) \geq P(\neg S_1 \wedge S_2)$, $P(S_1 \wedge S_2) \leq P(\neg S_1 \wedge \neg S_2)$, $P(S_1 \wedge \neg S_2) \geq P(\neg S_1 \wedge S_2)$, $P(S_1 \wedge \neg S_2) \leq P(\neg S_1 \wedge \neg S_2)$ and $P(\neg S_1 \wedge S_2) \leq P(\neg S_1 \wedge \neg S_2)$.

The probabilistic logic program for computing $P(S_a)$ was write as:

$$
\begin{array}{l}
\min / \max \quad P(R_a) \\
s.t \\
\\
\begin{array}{ll}
P(\neg S_1 \wedge S_2) \leq P(\neg S_1 \wedge \neg S_2) & P(S_2|S_0) \geq 0.82 \\
P(S_1 \wedge S_2) \geq P(S_1 \wedge \neg S_2) & P(S_1|S_0) \geq 0.51 \\
P(S_1 \wedge S_2) \geq P(\neg S_1 \wedge S_2) & P(S_0) = 0.0008 \\
P(S_1 \wedge \neg S_2) \geq P(\neg S_1 \wedge S_2) & P(S_1) \geq 0.4 \\
P(S_1 \wedge \neg S_2) \geq P(\neg S_1 \wedge \neg S_2) & P(S_2) \geq 0.33 \\
P(\neg S_1 \wedge S_2) \geq P(\neg S_1 \wedge \neg S_2) &
\end{array}
\end{array}
\tag{15}
$$

Program 15 was converted into a linear program and solved with the revised simplex algorithm. It resulted in $P(R_a) \in [0.5, 0.8]$ and self-information $I(R_a) \in [0.32, 1]$. The obtained probability interval was combined with the likelihoods (see Gay and Boullé (2013)) of R_a to calculate the interestingness level: $level(R_a) \in [0.523, 0.527]$. For R_b and R_c, an analogous procedure resulted in $P(R_b) = [0.66, 1]$, $P(R_c) = [0.66, 1]$, $I(R_b) \in [0, 0.59]$, $I(R_c) \in [0, 0.59]$, $level(R_b) \in [0.511, 0.514]$, and $level(R_c) \in [0.384, 0.387]$.

The results show that, given the knowledge base, rules R_b and R_c are relatively expected (self-information upper bound less than 1). For the rule R_a, self-information is not too revealing about its agreement (expectedness) with the prior knowledge. The lower bound of the self-information of R_a, 1, indicates that knowledge base does not allow to say that is un/expected. The level measure indicates that the rules R_a, R_b and R_c explain data with at a lower cost than the default rule (scores were greater than zero).

Continuing with the example, the data analyst obtained information that could influence the expectations with respect to the first rule. A statistical report states that the mean values of *concave points n1* and *perimeter n3* are bounded by the intervals [75.22, 138.8] and [0.05; 0.08], and their variances pertain to the intervals [28, 34] and [0.032, 0.044]. The observed correlation was 0.85. After entering that information in Program 15, the probability interval of R_a is updated to $P(R_a) \in [0.64, 0.8]$. The updating of self-information and level produces the intervals $I(R_a) \in [0.32, 0.64]$ and $level(R_a) = [0.526, 0.527]$. The new value of $I(R_a)$ provides evidence that suggests that R_a is in line with the domain knowledge.

5 Discussion

Probabilistic reasoning has been widely used in the development of tools for interestingness analysis (Bie; 2011)(Hahsler and Hornik; 2008)(McGarry; 2005)(Silberschatz and Tuzhilin; 1996). In this context, the approach presented here bears similarities to those described by Jaroszewicz et al. (2009) and Malhas and Aghbari (2009). Those authors proposed two approaches for subjective interestingness analysis of association rules. To do so, they used the formalism of Bayesian networks to represent the domain knowledge and explored Bayesian networks inference facilities to implement procedures that allow to compute interestingness scores based on entropy.

Apart from addressing classification rules, the proposed approach differs from the works of Jaroszewicz et al. (2009) and Malhas and Aghbari (2009) by its use of probabilistic logic programming to reason about domain knowledge. It organizes the knowledge base as a set of sentences defined on propositions built on the domain objects. The sentences express local relationships involving marginal, conditional, and bivariate probabilistic statements (although it is possible to formulate more complex statements).

By exploring a probabilistic logic-based schema, the proposed approach can be used even if the information elicited from experts, reports, descriptive statistics, or correlation data is not sufficient to specify a complete probabilistic model. As observed by Barbaros et al. (2014), Berleant and Jianzhong (2004) and Nitti et al. (2016), it is often the case. Additionally, the inference procedure enables reasoning with uncertain and incomplete knowledge (Haenni et al.; 2013) (Kern-Isberner et al.; 2011), qualitative probabilities (Ognjanovi et al.; 2008), and imprecise beliefs (Hansen et al.; 2000).

Finally, it must be noted that: (a) probabilistic logic inference is a time-consuming task (Cozman and Maua; 2017) (Hansen and Perron; 2008); (b) depending on the application, rule miners can generate excessive patterns (Balaji and Rao; 2013); and (c) it is desirable that interestingness analysis algorithms run rapidly. It is likely that these requirements will cause difficulties when applying the proposed approach in complex domains, mainly if it is necessary to process on a very large knowledge base. However, a further scrutiny demonstrates that rule mining algorithms generally implement a rule pruning strategy. Therefore, generally, the mined rules will not have numerous terms and, in this manner, the related inference problems are likely to have few terms and could be solved quickly (Hansen and Perron; 2008) (Cozman et al.; 2006) (Desrosiers

[3] See http://www.cdc.gov/mmwr/preview/mmwrhtml/00043942.htm.

[4] For practical reasons, all the probabilities were estimated from a random sample extracted from the original data set.

and Lubbecke; 2005) (Hansen et al.; 2000).

6 Conclusion

This work presented an approach for subjective interestingness analysis of classification rules. By using propositional probabilistic logic as a knowledge representation scheme, it allows the codification of the domain knowledge acquired from experts and information extracted from statistical reports in a unified way. It also allows the exploration of the commonly used inference algorithms in order to compute probability-based interestingness measures. Another advantage is that it is possible to carry out valid computations even if the available knowledge is uncertain or incomplete and the elicited beliefs are imprecise.

In the near future, we intend to extend the proposed approach in order to integrate independence assumptions into the reasoning. We also intend to employ the described approach for interestingness analysis of the association rules.

Acknowledgements

We thank CAPES and Fundação Araucária for their partial support.

References

Balaji, B. V. and Rao, V. V. (2013). Improved classification based association rule mining, *Intl. Journal of Advanced Research in Computer and Communication Engineering* **2**(5): 2211–2221.

Barbaros, Y., Perkins, Z., Rasmussen, T., Tai, N. and Marsh, D. (2014). Combining data and meta-analysis to build bayesian networks for clinical decision support, *Journal of Biomedical Informatics* **52**: 373–385.

Berleant, D., Ceberio, M., Xiang, G. and Kreinovich, V. (2007). Towards adding probabilities and correlations to interval computations, *Int. J. Approx. Reasoning* **46**(3): 499–510.

Berleant, D. and Jianzhong, Z. (2004). Using pearson correlation to improve envelopes around the distributions of functions., *Reliable Computing* **10**(2): 139–161.

Bie, T. D. (2011). Maximum entropy models and subjective interestingness: an application to tiles in binary databases., *Data Mining and Knowledge Discovery* **23**(3): 407–446.

Cohen, W. W. (1995). Fast effective rule induction, *In Proceedings of the Twelfth International Conference on Machine Learning*, Morgan Kaufmann, pp. 115–123.

Cozman, F. G., de Campos, C. and da Rocha, J. (2006). Probabilistic logic with strong independence, *in* J. Sichman, H. Coelho and S. Rezende (eds), *Advances in Artificial Intelligence, IBERAMIA-SBIA 2006, Brazilian Symposium on Artificial Intelligence*, Vol. 4140 of *Lecture Notes in Computer Science*, pp. 612–621.

Cozman, F. and Maua, D. (2017). On the semantics and complexity ofprobabilistic logic programs, *Journal of Artificial Intelligence Research* **60**: 221–262.

Desrosiers, J. and Lubbecke, M. (2005). *Column Generation*, Springer, Boston, USA, chapter A Primer in Column Generation, pp. 1–32.

Egho, E., Gay, D., Boullé, M., Voisine, N. and Clérot, F. (2015). A parameter-free approach for mining robust sequential classification rules., *in* C. Aggarwal, Z.-H. Zhou, A. Tuzhilin, H. Xiong and X. Wu (eds), *International Conference on Data Mining*, IEEE Computer Society, pp. 745–750.

Fürnkranz, J., Gamberger, D. and Lavrač, N. (2014). *Foundations of Rule Learning*, Springer Publishing Company, Incorporated.

Garthwaite, P. H., Kadane, J. B. and O'Hagan, A. (2005). Statistical methods for eliciting probability distributions, *Journal of the American Statistical Association* **100**: 680–701.

Gay, D. and Boullé, M. (2013). A bayesian criterion for evaluating the robustness of classification rules in binary data sets, *Advances in Knowledge Discovery and Management*, pp. 3–21.

Geng, L. and Hamilton, H. J. (2006). Interestingness measures for data mining: A survey, *ACM Comput. Surv.* **38**(3).

Haenni, R., Romeijn, J.-W., Wheeler, G. and Williamson, J. (2013). *Probabilistic Logics and Probabilistic Networks*, Springer Publishing Company, Incorporated.

Hahsler, M. and Hornik, K. (2008). New probabilistic interest measures for association rules, *CoRR* **abs/0803.0966**.

Hamilton, A. (1988). *Logic for Mathematicians*, Cambridge University Press.

Han, J. (2005). *Data Mining: Concepts and Techniques*, Morgan Kaufmann Publishers Inc., San Francisco, CA, USA.

Hansen, P., Jaumard, B., de Aragão, M. P., Chauny, F. and Perron, S. (2000). Probabilistic satisfiability with imprecise probabilities, *International Journal of Approximate Reasoning* **24**(2–3): 171 – 189.

Hansen, P. and Perron, S. (2008). Merging the local and global approaches to probabilistic satisfiability, *International Journal of Approximate Reasoning* **47**(2): 125 – 140.

Hooker, J. N. (1992). Mathematical programming methods for reasoning under uncertainty, *in* W. Gaul, A. Bachem, W. Habenicht, W. Runge and W. Stahl (eds), *Operations Research 1991*, Vol. 1991 of *Operations Research Proceedings 1991*, Springer Berlin Heidelberg, pp. 23–34.

Jaroszewicz, S., Scheffer, T. and Simovici, D. A. (2009). Scalable pattern mining with bayesian networks as background knowledge., *Data Min. Knowl. Discov.* **18**(1): 56–100.

Kern-Isberner, G., Beierle, C., Finthammer, M. and Thimm, M. (2011). *Probabilistic Logics in Expert Systems: Approaches, Implementations, and Applications*, Springer Berlin Heidelberg, Berlin, Heidelberg, pp. 27–46.

Leeuwen, M., Bie, T., Spyropoulou, E. and Mesnage, C. (2016). Subjective interestingness of subgraph patterns, *Machine Learning* **105**(1): 41–75.

Levi, I. (1980). *The Enterprise of Knowledge*, MIT Press, Cambridge.

Malhas, R. and Aghbari, Z. A. (2009). Interestingness filtering engine: Mining bayesian networks for interesting patterns, *Expert Syst. Appl.* **36**(3): 5137–5145.

McGarry, K. (2005). A survey of interestingness measures for knowledge discovery, *Knowl. Eng. Rev.* **20**(1): 39–61.

Nitti, D., Ravkic, I., Davis, J. and Raedt, L. D. (2016). Learning the structure of dynamic hybrid relational models, *22nd European Conference on Artificial Intelligence, Prestigious Applications of Artificial Intelligence 2016*, The Hague, Netherlands, pp. 1283–1290.

Ognjanovi, Z., Perovi, A. and kovi, M. R. (2008). An axiomatizatin of qualitative probability, *Acta Polytechnica Hungarica* **1**(5): 105–110.

O'Hagan, A., Buck, C., Daneshkhah, A., Eiser, J., Garthwaite, P., Jenkinson, D., Oakley, J. and Rakow, T. (2006). *Uncertain judgements: eliciting experts' probabilities*, John Wiley and Sons, Chichester.

Russell, S. and Norvig, P. (2010). *Artificial Intelligence: A modern approach*, 3a edn, Prentice Hall, Upper Saddle River.

Silberschatz, A. and Tuzhilin, A. (1996). What makes patterns interesting in knowledge discovery systems, *IEEE Transactions on Knowledge and Data Engineering* **8**(6): 970–974.

Sivia, D. S. and Skilling, J. (2006). *Data Analysis, A Bayesian Tutorial*, 2nd edn, Oxford University Press, New York.

van der Gaag, L., Renooij, S., Witteman, C., Aleman, B. and Taal, B. (2013). How to elicit many probabilities, *CoRR* **abs/1301.6745**.

Vashishtha, J., Kumar, D., Ratnoo, S. and Kundu, K. (2011). Article: Mining comprehensible and interesting rules: A genetic algorithm approach, *International Journal of Computer Applications* **31**(1): 39–47.

Walley, P. (1991). *Statistical Reasoning with Imprecise Probabilities*, Monographs on Statistics and Applied Probability, Chapman and Hall, London.

Wolberg., W., Street, W. and Mangasarian, O. (1994). Machine learning techniques to diagnose breast cancer from image-processed nuclear features of fine needle aspirates., *Cancer Lett* **77**(2-3): 163–71. **URL:** *http://www.biomedsearch.com/nih/Machine-learning-techniques-to-diagnose/8168063.html*