A probabilistic logic-based approach for subjective interestingness analysis

José Carlos Ferreira da Rocha, Alaine Margarete Guimarães and Valter Luis Estevam Jr

Abstract— **This paper describes 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 which encodes information incoming from experts and statistical data. Computation of interestingness scores is supported by a reasoning procedure that uses linear programming to calculate the probabilities of interest. An application example illustrates the utilization of the described approach in evaluating classification rules on UCI Wisconsin Cancer data set. As it is shown this scheme provides a mechanism to estimate probability based subjective interestingness scores.**

I. Introduction

Searching patterns in databases has been a useful strategy for acquiring the knowledge that is required to perform tasks involving decision making and problem-solving in areas such as medicine, agriculture, biology, environmental research and many others. In this context, knowledge discovery in data bases (KDD), the field of computer science, intends to develop the theoretical basis and computational frameworks to transform raw data into useful and comprehensive information. In practice, KDD process can be abstracted in three basic steps: preprocessing data to make it ready for analysis, running data mining algorithms to find out relationships among the variables of the application domain and then, evaluating how interesting are the detected patterns.

A common application in KDD is mining classification rules that aiming to finding logical implications which relate the features of an object to a label that informs its category (or class) by examining cases in a data set. The main criteria for evaluating rule mining results is the accuracy, the overall correctness of the model in predicting object class in a test data set. Basically, if the discovered rules are sufficiently accurate it is possible to use them to classify new object instances given the evidence.

However, depending on application objectives, accuracy is not the only relevant criterion in classification rule mining. In some situations, it might also be important to estimate the usefulness, consistency or novelty of the learned patterns in the light of domain knowledge. It is the aim of subjective interestingness analysis, a KDD procedure that explores domain knowledge to calculate scores that try to quantify how much a pattern meets a criterion of interest. For example, by verifying if the obtained rules are consistent with the background knowledge [1], [2]. - that is, if they are expected given the available information.

To perform this task, data analysts usually make use of a knowledge-based system which provides the functionalities required to compare the data mining results to previous expectations. Such approach demands the implementation of a knowledge base that encodes all relevant information and thus, involves carrying out a knowledge engineering process to acquire the background information and then writing the elicited sentences using an appropriated formalism. However, domain knowledge is, frequently, incomplete and imprecise and, in this case, the selected formalism must provide the mechanism to proceed reasoning under uncertainty.

Probability theory has been widely employed to represent uncertain beliefs about the statements in a knowledge base and perform inferences about expectations. However, as observed by Walley [3] [4], domain experts do not always feel comfortable to assign exact probabilities to the sentences of a knowledge base. In other terms, sometimes it can be the case that the information provided by the experts is not enough to the elicitation of point probability estimates. Since this situation can also arise during the development of a knowledge base for subjective interestingness analysis, the use formalisms which allow to handle with imprecise probability statements can be very useful in practice.

Considering it, this work presents an approach that explores the formalism of probabilistic logic to encode the prior knowledge and uses a linear programming-based procedure to compute interestingness scores for classification rules. The main objective underlying the proposal is to provide a basic framework which allows to integrate domain knowledge from different sources and also to execute the needed inferences. In particular, the proposed approach is able to deal with information from probabilistic distributions defined on the attributes which appear in a classification rule, statistics about correlation data and imprecise beliefs elicited from experts.

The paper is organized as follows. Section 2 brings a review on classification rule mining and an interestingness measure called level, proposed by Gay and M. Boullé [5]. This section also describes some concepts in probabilistic logic. Section 3 presents the proposed approach. Section 4 illustrates the utilization of the described approach with a simple application in which the level of interest of classification rules generated by the JRIP algorithm on the UCI Breast Cancer Data Set is calculated. Section 5 discusses the main issues related to the employment of the proposed approach in interestingness analysis. The last section presents the final

J.C.F. da Rocha is with the Computational Intelligence Lab, UEPG, Ponta Grossa, PR 84030-900, Brazil (email: jrocha@uepg.br).

A.M.Guimarães is with the Infoagro Lab, UEPG, Ponta Grossa, PR 84030-900, Brazil (email: alainemg@hotmail.com).

Valter L. Estevam Jr is with Department of Informatics at IFC, Videira, SC 84030-900, Brazil.

remarks of this study.

II. Background review

Let $\mathbf{D}$ be a multivariate data set with m instances, n descriptor attributes and a target attribute. Denote the descriptors by $X_1 ..., X_n$ and the target or class by C. Each attribute X_j symbolizes some feature of objects to be classified and its domain[1] is denoted as Ω_j. The target attribute is a category label.

Classification rule mining aims at discovering implication patterns that can be used to classify objects into given categories based on their features [6]. It is the task of inducing, from $\mathbf{D}$, logical expressions of the form $F_1 \wedge F_2 \cdots \wedge F_t \rightarrow C$. In this work, each antecedent F_i of a rule is assumed to be a relational expression $X_j \odot x_{j,k}$ where $x_{j,k} \in \Omega_j$, $\odot$ is an operator from the set $\{<, >, \leq, \geq, =\}$ and $t \in \mathbb{N}^+$ with $t \leq n$. The consequent can be viewed as a proposition labeling an object as a member of a particular class. There is a finite number of classes.

As it does for other data mining procedures used in KDD, it is necessary to verify whether the rule mining results meet the application requirements. In this kind of analysis, called interestingness analysis, some numerical scores, the interestingness measures, are computed for each discovered pattern. Good interestingness measures should indicate how much a pattern complies with KDD goals by identifying "valid, novel, potentially useful and ultimately understandable information in data" [7]. As there are many different interestingness scores in literature it is necessary to choose a suitable one for each application. Afterward, the selected score can be used to filter or rank the most promising results.

Since subjective interestingness analysis explores domain knowledge and user preferences to rank the discoveries, its realization requires building a knowledge base as well as providing the associated reasoning procedures. Furthermore, its implementation often faces some issues related to the development of knowledge-based systems. Two of them are how to deal with uncertain and incomplete knowledge [8], [9]. Probability theory has been a ubiquitous tool to handle these conditions and considering it, D. Gay and M. Boullé [5] proposed an interestingness score, named level, which is defined as follows:

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

In this expression, $c(R) = -log(P(\mathbf{D}|R)) - log(P(R))$ is said to be the cost of the rule and $c(R_0)$ is the cost of a default rule - it is computed from class frequencies in the data set. As it can be seen, $level(R)$ is a posterior probability-based score and as such evaluates, simultaneously, whether a rule fits to data and agrees with prior knowledge. If $level(R) \leq 0$, the rule is less than or as probable as the default rule and it is said uninteresting. If $0 < level(R) < 1$, R is more probable than the default rule and it is more interesting as its level approaches 1. If $level(R) = 1$, the rule fits the observations and prior beliefs exactly.

A point to be noted here is that eliciting probabilities from experts is difficult activity [10]. So, if previous data or literature provides information that can be relevant for interestingness analysis (probability distributions, descriptive statistics, correlation), it should be integrated into the knowledge base whenever it is possible. Another point is that eliciting probabilities from experts or literature does not always manage to obtain exact probability assignments. In particular, in some domains, it can be the case that all available information is formed by imprecise probabilities [3] or qualitative beliefs [11]. From this follows that it can be useful to select formalisms which are is able to deal with numeric and interval-valued probabilities and qualitative probabilities [12] [13].

A. Propositional probabilistic logic

Probabilistic logic provides a formalism that extends propositional logic for dealing with uncertain knowledge [14]. As propositional logic, probabilistic logic also uses propositional variables to represent categorical statements. Propositional variables, or atomic formulas, can assume one of two states, *true* or *false*, and can be combined in order to form a compound formula. A compound formula describes a complex proposition and is obtained by connecting atomic or other compound formulas using the logical operators. In this work, atomic formulas are denoted by lower case letters as $p \ldots, q$, compound formulas are denoted by capital letters $A, B \ldots, C$ and logic operators ($\wedge, \vee$ and $\neg$) has the usual semantics [15]

Let S_i be a formula, atomic or not. Probabilistic logic assumes that the agent's belief in S_i can be represented by a probability assignment $P(S_i) = \pi_i$, with $\pi_i \in [0, 1]$. If beliefs are imprecise, they can be expressed by inequalities as $P(S_i) \geq \pi_i$ or $P(S_i) \leq \pi_i$ or by interval probability statements 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 for π_i. Furthermore, exact conditional probability statements expressing the expectation in a sentence S_i given the event S_j can be written as $P(S_i|S_j) = \pi_{i,j}$, $P(S_i|S_j) \geq= \pi_{i,j}$. As before, imprecise conditional beliefs can be represented as inequalities.

A probabilistic logic knowledge base is said consistent if its assignments agree with probability theory axioms. So, if $\mathcal{M}$ denotes all possible truth assignments[2] on the variables of a consistent knowledge base, then:

$$P(S_i) = \sum_{w:\mathcal{M}(S_i,w)} P(w)\,. \tag{2}$$

where $P(w)$ is the probability of a truth assignment.

A *probabilistic logic inference* [16] aims at computing $\underline{P}(S)$ and $\overline{P}(S)$, the lower and upper probability of a sentence S given the constraints defined by a knowledge base with assessments $P(S_*) = \pi_*$, $P(S_*) \leq \pi_*$ $P(S_*) \geq$

[1] In this work we assume a variable can be categorical, discrete or continuous.

[2] A truth assignment is a vector assigning value either true or false to each propositional variable of an expression to the constants *true* or *false*.

π_*. The resultant interval $[\underline{P}(S), \overline{P}(S)]$ must be consistent with the assessments about every sentence S_* and with the probability theory axioms. In this scheme, the given initial assessments compose a knowledge base which encodes the prior information as also any relevant evidence.

Inference can be carried out by linear programming. Basically, let $\mathbf{S} = \{S_1, \ldots, S_m\}$ be a set of propositional sentences associated to a collection of probability assignments $\Pi = \{\pi_1 \ldots \pi_m\}$ and let S be the sentence of interest whose probability is unknown. Let $\mathbf{a}_i$ be a row vector so that the j^{th}element of $\mathbf{a}_i$ is 1 if S_i is true in w_j, the j^{th} truth assignment, otherwise it is zero. Let also $\mathbf{p} = (p_1 \ldots, p_{2^n})^T$ be a vector with the probability of every truth assignment defined on the atomic sentences in $\mathbf{S} \cup \{S\}$. From Equation 2 it is easy to see that $P(S_i) = \mathbf{a}_i^T \mathbf{p}$. A similar vector $\mathbf{a}$ can also be defined for S and the linear programming related to $P(S)$ can be written as:

$$\begin{array}{ll} \min / \max & \mathbf{a}^T \mathbf{p} \\ s.t & A_{m \times 2^n} \mathbf{p} = \pi \\ & p_i \geq 0, \ i = 1..2^n \\ & \mathbf{1p} = 1 \end{array}$$

In this program, $\mathbf{a}_1 \ldots \mathbf{a}_m$ are the rows of matrix A.

III. Probabilistic logic programming and interestingness analysis

Let $F_1 \wedge F_2 \cdots \wedge F_t \rightarrow C$ be a classification rule and S a logical variable which is equivalent to that. Let S_i be a propositional variable that stands for the antecedent F_i. As before, F_i represents a statement in the form $X_j = x_{j,k}$, $X_j \geq x_{j,k}$ or $X_j \leq x_{j,k}$. The consequent C is a sentence S_0 indicating a class labeling.

In the proposed approach, the first step for the subjective interestingness analysis of a rule is to elicit the marginal probabilities of rule antecedents. In this work, it is assumed that this information can be elicited from experts [10], obtained from previous data analysis results [17], learned from domain literature [18] or derived by meta-analysis [11].

Alternatively, the probability of each rule component could be calculated from the empirical or theoretical marginal densities related to its respective attribute [19]. So, let it be an initial assumption that $p(X_j)$ is known for every $X_j \in \mathbf{X}$. In this case, it is always possible to determine the value of π_i for every S_i and then to generate the constraint $P(S_i) = \pi_j$. Given that, lower and upper bounds for $P(S)$ can be calculated by solving the next probabilistic program:

$$\begin{array}{ll} \min / \max & P(S) \\ s.t. & P(S_1) = \pi_1 \\ & \ldots \\ & P(S_t) = \pi_t \end{array} \tag{3}$$

Now, a problem can arise whether $p(X_j)$ is unknown or uncertain. Nevertheless, the proposed approach can be extended in many ways to deal with this issue. For example, a first course of action would be, as stated before, to explore domain experts knowledge or literature. A second choice would be to use ignorance priors [20]. A third alternative would be to employ the imprecise probability theory [21], [3] to obtain a probability interval $[\underline{\pi}_i, \overline{\pi}_i]$ which could be used to define the expression $\underline{\pi}_i \leq P(S_i) \leq \overline{\pi}_i$. In any case, the resultant equations and inequalities could be appended to Program (3) in order to proceed the inferences.

After encoding relevant information in Program (3), it can be converted into a linear program for later solving. Example 1 illustrates the process described here.

> Example 1: Let X_1 and X_2 be two normally distributed variables so that $X_1 \sim N(1; 0.1)$ and $X_2 \sim N(4; 1)$ and let $C = c_1$ be a class assignment whose prevalence is greater or equal to 0.6. Given a rule $S \equiv (X_1 \leq \wedge X_2 \leq 5 \rightarrow C = c_1)$, it is possible to use the previous information to build a probabilistic logic program for $P(S)$. In that program, $P(S_1) = 0.16$, $P(S_2) = 0,84$ and $P(S_0) \geq 0,6$. The upper and lower bounds for $P(S)$ are obtained by solving the next:
>
> $$\begin{array}{ll} \min / \max & \mathbf{ap} \\ s.t & \mathcal{A} \times \mathbf{p} = \Pi \\ & \mathbf{1p} \\ & p_i \geq 0, i = 1..8. \end{array}$$
>
> In this program, $\mathcal{A} = \begin{bmatrix} \mathbf{a}_0 \\ \mathbf{a}_1 \\ \mathbf{a}_2 \end{bmatrix}$, $\mathbf{p} = \begin{pmatrix} p_1 \\ \ldots \\ p_8 \end{pmatrix}$ and $\Pi = \begin{pmatrix} 0.65 \\ 0.16 \\ 0.84 \end{pmatrix}$. The rows of $\mathcal{A}$ and the objective function 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 $\mathbf{a} = (1,0,1,1,1,1,1,1)$.

It can also be the case that domain experts do not feel comfortable to assign bounds to the probabilities of some sentences but have information that allows to ascertain comparative probability statements. For example, let Q_1, Q_2 and Q_3 be three sentences defined on $S_1 \ldots, S_t$ so that experts know that: (a) Q_1 is as probable as or more probable than Q_2 and (b) Q_3 is as probable as or more probable than Q_1. This kind of statement can be incorporated to Program (3) as $P(Q_1) \geq P(Q_2)$ and $P(Q_3) \geq P(Q_1)$.

More formally, if it is known that $P(Q_1) \leq P(Q_2)$, $P(Q_1) \geq P(Q_2)$ or $P(Q_1) = P(Q_2)$, it is possible to use that qualitative information to generate expressions, on the form $P(R_1) - P(R_2) \leq 0$, $P(R_1) - P(R_2) = 0$ or $P(R_1) - P(R_2) \geq 0$, respectively. As before, such constraints can be rewritten using a vectorial notation by doing $\mathbf{b}_{1,2} = \mathbf{b}_1 - \mathbf{b}_2 \odot 0$. $\mathbf{b}_1$ and $\mathbf{b}_2$ are the row vectors relative to $P(Q_1)$ and $P(Q_2)$. Given a set of those constraints, grouped into a system $\mathcal{B} \times \mathbf{p} \odot \varpi$, they can be appended to the inference

program as follows:

$$\begin{array}{ll} \min / \max & \mathbf{c}^T\mathbf{p} \\ s.t & \left[\begin{array}{c}\mathcal{A}\\ \mathcal{B}\end{array}\right] \times \mathbf{p} \odot \left[\begin{array}{c}\varpi\\ \mathbf{0}\end{array}\right] \\ & \mathbf{1p} \\ & p_i \geq 0, i = 1..2^t. \end{array} \quad (4)$$

As it can be deduced from the expression above, by encoding the information provided by comparative probabilities into the linear program, it is possible to reduce the solution space of the optimization problem. So, it can contribute to obtaining tighter bounds for $P(S)$ and derived interestingness scores.

A. Dealing with correlation data

Berleant and Jianzhong [22] and Berleant *et al* [23] present a procedure that allows the calculation of envelopes for joint probabilities from Pearson's correlation coefficient and marginal data. This section explores that procedure to draw out additional probabilistic constraints for interestingness analysis.

Initially, let X_i and X_j be two distinct continuous attributes, with known densities $p(X_i)$ and $p(X_j)$, and linear correlation r. Discretization of X_i and X_j into n_1 and n_2 bins introduces two discrete variables, Z and Y whose sample spaces are $\Omega_z = \{z_1 \dots z_{n_1}\}$ and $\Omega_y = \{y_1 \dots y_{n_2}\}$. In addition, let $p(Z)$ and $p(Y)$ be the marginal distributions of these new variables so that their entries are obtained from $p(X_i)$ and $p(X_j)$ by doing $P(z_k) = P\big(\underline{x}_{i,k} < X_i \leq \overline{x}_{i,k}\big)$ and $P(y_l) = P\big(\underline{x}_{j,l} < X_j \leq \overline{x}_{j,l}\big)$. 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).

Given $S_i \equiv (X_i \geq x_i)$ and $S_j \equiv (X_j \geq x_j)$, two antecedents of a classification rule, if it is assumed that Z has a value z_a so that $\underline{x}_{i,a} = x_i$ and Y has a value y_b where $\underline{x}_{j,b} = x_j$, $P(S_i)$ and $P(S_j)$ can be easily written in terms of $p(Z)$ and $p(Y)$. Moreover, marginalization of $p(Z,Y)$ produces:

$$\begin{array}{l} P(S_i) = \sum_{k=z_a}^{z_{n_1}} \sum_{l=1}^{n_2} P(Z = z_k \wedge Y = y_l) = \pi_i \\ P(S_j) = \sum_{l=y_b}^{k_{n_2}} \sum_{k=1}^{n_1} P(Z = z_k \wedge Y = y_l) = \pi_j \end{array} \quad (5)$$

Similarly, $P(S_i \wedge S_j)$ can be formulated in terms of $p(Z,Y)$ by doing:

$$P(S_i \wedge S_j) = \sum_{t_* \in \mathbf{t}} P(Z = z_{t_*} \wedge Y = y_{t_*}) = \pi_{i\wedge j}. \quad (6)$$

In this expression, $\pi_{i\wedge j}$ denotes the unknown value $P(S_i \wedge S_j)$ and $\mathbf{t}$ is a vector of pairs of indexes so that, for all $t_* = (k,l) \in \mathbf{t}$, the intervals represented by z_k and y_l agree with the condition symbolized by $S_i \wedge S_j$. As before, Equations 5 and 6 can be represented in a vectorial form and appended to the Program (4).

The point here is that previous equations relate the joint distribution of $p(X_i, Xj)$ to $P(S_i \wedge S_j)$, $P(S_i)$ and $P(S_j)$ - all of them are relevant to compute the probability of the classification rule under analysis. The problem is that integrating that information into described approach depends on estimate or bound the value of $\pi_{i\wedge j}$. It can be done by exploring Equations 7 and 8:

$$\sum_{k,l}^{n_1,n_2} \overline{z_k y_l} P(Z = z_k \wedge Y = y_l) \geq \underline{\mu_i \mu_j} + r\sqrt{\underline{\sigma_i^2 \sigma_j^2}} \quad (7)$$

$$\sum_{k,l}^{n_1,n_2} \underline{z_k y_l} P(Z = z_k \wedge Y = y_l) \leq \overline{\mu_i \mu_j} + r\sqrt{\overline{\sigma_i^2 \sigma_j^2}} \quad (8)$$

As shown by Berleant and Jianzhong [22] those equations can be used to calculate an outer envelope for $p(Z,Y)$ given correlation data and some statistical measures. Basically, beyond the correlation of X_i and X_j, their utilization requires that outer bounds on the expected values ($\underline{\mu_*}$ and $\overline{\mu_*}$) and variances ($\underline{\sigma_*^2}$ and $\overline{\sigma_*^2}$) be known.

Equations 5, 6, 7 and 8 can be grouped to form a linear system $\mathcal{D}$ that also stores the constraints implied by the probability theory. Appending $\mathcal{D}$ to Program (4) can be useful for two reasons. Firstly, because it provides additional constraints to the optimization program and, from this, contributes to obtaining tighter intervals for $P(S)$. Secondly, because it also allows the collection of information in another way.

The last point to be discussed here is a note about the acquisition of $\underline{\mu_j}$, $\overline{\mu_j}$, $\underline{\sigma_i^2}$, $\underline{\sigma_j^2}$, $\overline{\sigma_i^2}$, $\overline{\sigma_j^2}$. As proposed by Berleant and Jianzhong (2004) and Berleant *et al* (2007), this work assumes that those limits are entered by the analyst or calculated by interval optimization upon $P(Z)$ and $P(Y)$.

B. Evaluating interestingness

The described approach assumes that interestingness analysis is performed as a post-processing routine. That is, interestingness analysis is performed after the data mining step and it aims at sorting or filtering the mined rules according to their interestingness scores.

Given that, calculating the level of interestingness of a rule starts by computing $P(S)$ with the model previously described and then employing that result to determine the numerator, $c(S) = -\log(P(S)) - \log(P(\mathbf{D}|S))$ in Equation 1. If $P(S)$ is determined exactly, $c(S)$ can be calculated directly by using Equation 1. Otherwise, if the result is an interval $[\underline{P}(S), \overline{P}(S)]$, Equation 1 can be used to derive an interval for $c(S)$. In this case, since $\underline{c}(S) = -log(\overline{P}(S)) - log(P(\mathbf{D}|S))$ and $\overline{c}(S) = -log(\underline{P}(S)) - log(P(\mathbf{D}|S))$ are the limits of such interval, the minimum and maximum of level score are given by :

$$\begin{array}{l} \underline{level}(S) = 1 - \frac{\overline{c}(S)}{c(S_0)} \\ \overline{level}(S) = 1 - \frac{\underline{c}(S)}{c(S_0)} \end{array} \quad (9)$$

After obtaining the interval for $level(S)$, the interestingness analysis continues by inspecting its lower and upper bounds. if $\underline{level}(S)$ is greater than 0, it means that the rule appears to be interesting given prior knowledge as also effective in describing data, even if it was computed on the lower bound for $P(S)$. On the other hand, $\overline{level}(S) < 0$ is

an indicative that, in the light of 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 direct conclusion can be drawn.

IV. An application example

This section shows an application that uses the proposed approach to calculate the level of interestingness of classification rules induced by JRIP algorithm [24], [25] for the Breast Cancer Wisconsin Data Set [26]. The 569 cases in data set were split into two partitions, the training data with $2/3$ of the instances and a test data set with the rest. The JRIP algorithm generated the following 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 attributes on the left side of these rules refer to some features of cellular nucleous. The right side indicates positive diagnostic of malignancy.

As prescribed by the proposed approach, the first rule was associated with a sentence $S_a \equiv S_1 \wedge S_2 \rightarrow S_0$ where S_1 and S_2 symbolize the conditions *concave points n1* $\geq$ *0.05182* and *perimeter n3* $\geq$ *113.9*, respectively. S_0 denotes the sentence *Diagnosis=malign*. The prior probabilities of S_1 and S_2 were set as $P(S_1) = \pi_1 = 0.41$ and $P(S_2) = \pi_2 = 0.36$ and the conditional probability $P(S_1|S_0)$ was also entered as $\pi_{1,0} = 0.82$. In the example it was assumed that those values were informed by an hypothetical expert[3]. The prevalence of breast cancer incidence[4] in US was defined as $P(S_0) = \pi_0 = 0.001$. Next, the following probabilistic logic program was written as:

$$\begin{array}{ll} \min / \max & P(S_a) \\ s.t & P(S_0) = 0.001 \\ & P(S_1) = 0.41 \\ & P(S_2) = 0.36 \\ & P(S_1|S_0) = 0.82 \end{array}$$

Similarly, Rule (b) was associated with a sentence rule $S_b \equiv S_3 \wedge S_4 \rightarrow S_0$ where S_3 and S_4 represent the propositions *concave points n1* $\geq$ *0.05839* and S_4 and *texture n3* $\geq$ *23.75*. The input probabilities were fixed as $P(S_3) = 0.34$, $P(S_4) = 0.58$ and $P(S_4|S_0) = 0.85$. Rule (c) was related to sentence $S_c \equiv S_5 \wedge S_6 \wedge S_7 \rightarrow S_0$ where S_5, S_6 and S_7 indicate *radius n3* $\geq$ *15.65*, *texture n3* $\geq$ *28.06* and *smoothness n3* $\geq$ *0.1094*, respectively. The input probabilities were $\pi_5 = 0.4987$, $\pi_6 = 0.3398$, $\pi_7 = 0.8452$ and $\pi_{6,0} = 0.4425$.

After that, the revised simplex algorithm was used to calculate intervals for $P(S_a)$, $P(S_b)$ and $P(S_c)$, the obtained results were $[0.64, 1]$, $[0.66, 1]$ and $[0.66, 1]$, respectively. Next, the calculated probability intervals were combined with the likelihoods (see [5]) of S_a, S_b and S_c in order to calculate lower and upper bounds for the level score. The results were $level(S_a) \in [0.65, 0.75]$, $level(S_b) \in [0.026, 0.026]$ and $level(S_c) \in [0.023, 0.023]$. These results indicate that only the first rule seems to have a considerable degree of interest when confronting data and background knowledge.

In the sequence, it is supposed that the analyst has two additional pieces of information he wants to take into account when evaluating the first rule. The first one informs a qualitative constraint $P(S_1 \wedge S_0) \geq P(S_2 \wedge S_0)$. The second one, supplied by an expert, declares that the expected value of *concave points n1* and *perimeter n3* are bounded by the intervals $[75.22, 138.8]$ and $[0.05; 0.08]$ while their variances pertain to intervals $[28, 34]$ and $[0.032, 0.044]$.

By solving the linear program updated with this information, a new belief interval for S_a is $P(S_a) \in [0.99, 1]$. It makes that the lower bound for $level(S_a)$ be updated to 0.66.

V. Discussion

The probabilistic approach presented in this work provides a basic scheme to encode into a knowledge base information acquired from experts, literature or statistical reports aiming to make advances in the interestingness analysis of classification rules. In particular, it allows using information elicited from experts or descriptive statistical data to assess the marginal or joint probabilities for the propositions which compose a classification rule. Once this kind of information is often available in several domains [11] [22], the proposed approach can be useful in many situations.

Additionally, by exploring the probabilistic logic language, the presented approach implements two facilities which are inherent to that logic. A first facility stems from the fact that probabilistic logic reasoning is able to deal with uncertain and incomplete knowledge [27], [28], [29], qualitative probabilities [30] and imprecise beliefs [31] using the same inference engine - linear programming. A second one is that probabilistic logic modeling does not require the construction of a complete probabilistic model in order the encoding sentences involving many domain variables. This way, since the statements in the knowledge base comply with the axioms of probability theory [32], the declaration of any statement (with few or several terms) does not depend on prior statements.

Many other authors have used probabilistic reasoning methods in subjective interestingness analysis [33] [34][35] [36]. In particular, the present proposal bears some similarities with those described by Jaroszewicz and Simovici [37] and Malhas [8]. Those authors show schemes for subjective interestingness analysis of association rules where they apply the formalism of Bayesian networks [38] to represent the domain knowledge and proceed the computation of an interestingness score. That design choice allows them to take advantage from the expressiveness of the language and efficiency of reasoning algorithms provided by the Bayesian network formalism. However, Bayesian networks

[3]For practical reasons, $P(S_1)$, $P(S_2)$ and $P(S_1|S_0)$ were estimated from a random sample extracted from the original data set.

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

reasoning assumes that model probabilities are precise and the treatment of imprecise probabilities often demands the employing of extended formalisms [39] [40] [41]. Give that, the present approach can be viewed as an alternative for domains which are well represented by rule-based systems and the development team does not have time or resources to specify a complete probabilistic model.

Finally, it must be noted that: (a) probabilistic logic inference is a time-consuming task [16]; (b) depending on the application and the learning strategy, rule miners can generate too many patterns [42]; and (c) it is desirable that interestingness analysis tools have a pleasing time performance. So, given (a) and (b), it could be argued that the presented scheme would be ineffective to achieve (c) if it is necessary to reason on a very large knowledge base or if there are too many rule patterns to process. However, a further scrutiny shows that rule mining algorithms, usually, implement a kind of Occam's razor or strategy for rule pruning. So, for some applications, the mined rules will not have many components (propositions). In such cases, it is expected that the inference problem does not have too many elements to process and, therefore, it can be solved quickly. Additionally, empirical results described in Cozman, de Campos and da Rocha [43], Jaumard, Hansen and Aragão [14] and Hansen *et al* [31], shows that the use of column generation technique [44] allows to solve the linear programs related to probabilistic logic programs with a few dozens of variables and two hundred of sentences in less than a minute (few seconds). This time performance can be acceptable for a number of real world tasks.

VI. Conclusion

This work presented a propositional probabilistic logic-based approach for subjective interestingness analysis of classification rules. An advantage of the proposed approach is that it allows to integrate domain knowledge from experts, literature and statistical reports to compute probability based-interestingness measures. Another advantage is that it is possible to carry out valid probabilistic reasoning even if available knowledge is uncertain or incomplete and elicited beliefs are imprecise.

As a future work it is intended to extend the proposed approach to integrate independence assumptions into the inference step by combining probabilistic logics and graph based representations (see [45] and [43]). Another objective is to investigate the possibility of using the proposed scheme to validate or review prior beliefs, given the mined rules (similar to the soft belief analysis described by Silberschatz and Tuzhilin [36]). Finally, the development of an extended approach for dealing with interestingness analysis of association rules is also intended.

Acknowledgment

Thanks to CAPES, CNPq, Finep and Fundação Araucária for partially support this work.

References

[1] S. Chen and B. Liu, "Generating classification rules according to user's existing knowledge," in *First SIAM International Conference on Data Mining*, 2011, pp. 1–15.

[2] B. Liu, W. Hsu, and S. Chen, "Using general impressions to analyze discovered classification rules," in *3rd Intl. Conf. on Knowledge Discovery and Data Mining*, 1997, pp. 31–36.

[3] P. Walley, *Statistical Reasoning with Imprecise Probabilities*, ser. Monographs on Statistics and Applied Probability. London: Chapman and Hall, 1991.

[4] ——, "Measures of uncertainty in expert systems," *Artificial Intelligence*, vol. 83, no. 1, pp. 1–58, 1996.

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

[6] J. Vashishtha, D. Kumar, S. Ratnoo, and K. Kundu, "Article: Mining comprehensible and interesting rules: A genetic algorithm approach," *International Journal of Computer Applications*, vol. 31, no. 1, pp. 39–47, October 2011.

[7] U. Fayyad, G. Piatestky-Shapiro, and P. Smyth, "From data mining to knowledge discovery in databases," *AI Magazine*, vol. 17, no. 3, pp. 37–54, 1996.

[8] R. Malhas and Z. A. Aghbari, "Interestingness filtering engine: Mining bayesian networks for interesting patterns," *Expert Syst. Appl.*, vol. 36, no. 3, pp. 5137–5145, 2009.

[9] T. Yu, S. Simoff, and D. Stokes, "Incorporating prior domain knowledge into a kernel based feature selection algorithm," in *Advances in Knowledge Discovery and Data Mining*, ser. Lecture Notes in Computer Science, Z.-H. Zhou, H. Li, and Q. Yang, Eds. Springer Berlin Heidelberg, 2007, vol. 4426, pp. 1064–1071.

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

[11] Y. Barbaros, Z. Perkins, T. Rasmussen, N. Tai, and D. Marsh, "Combining data and meta-analysis to build bayesian networks for clinical decision support," *Journal of Biomedical Informatics*, p. in press, 2014.

[12] A. A. Silva and F. de Souza, "A protocol for the elicitation of imprecise probabilities." in *4th Intl. Symp. on Imprecise Probabilities: theory and applications*, F. Cozman, R. Nau, and T. Seidenfeld, Eds. SIPTA, 2005, pp. 315–321.

[13] L. G. de O. Silva and A. A. Filho, "A method for elicitation and combination of imprecise probabilities: A mathematical programming approach," in *2014 IEEE International Conference on Systems, Man, and Cybernetics*, 2014, pp. 619–624.

[14] B. Jaumard, P. Hansen, and M. D. Aragao, "Column Generation Methods for Probabilistic Logic," in *Integer Programming and Combinatorial Optimization*, 1990, pp. 313–331.

[15] A. Hamilton, *Logic for Mathematicians*. Cambridge University Press, 1988.

[16] P. Hansen and B. Jaumard, "Probabilistic satisfiability," École Polytechique de Montréal, Technical Report G-96-31, 1996.

[17] P. H. Garthwaite, J. B. Kadane, and A. O'Hagan, "Statistical methods for eliciting probability distributions," *Journal of the American Statistical Association*, vol. 100, pp. 680–701, 2005.

[18] L. van der Gaag, S. Renooij, C. Witteman, B. Aleman, and B. Taal, "How to elicit many probabilities," in *Proceedings of the 15th Conference on Uncertainty in Artificial Intelligence*, Stockholm, Sweden, 1999, pp. 647–654.

[19] D. S. Sivia, *Data Analysis, A Bayesian Tutorial*. New York: Oxford University Press, 1996.

[20] P. Gregory, *Bayesian Logical Data Analysis for the Physical Sciences*. New York, NY, USA: Cambridge University Press, 2005.

[21] I. Levi, *The Enterprise of Knowledge*. Cambridge: MIT Press, 1980.

[22] D. Berleant and Z. Jianzhong, "Using pearson correlation to improve envelopes around the distributions of functions." *Reliable Computing*, vol. 10, no. 2, pp. 139–161, 2004.

[23] D. Berleant, M. Ceberio, G. Xiang, and V. Kreinovich, "Towards adding probabilities and correlations to interval computations," *Int. J. Approx. Reasoning*, vol. 46, no. 3, pp. 499–510, 2007.

[24] W. W. Cohen, "Fast effective rule induction," in *In Proceedings of the Twelfth International Conference on Machine Learning*. Morgan Kaufmann, 1995, pp. 115–123.

[25] B. Martin, "Instance-based learning : Nearest neighbor with generalization," Tech. Rep., 1995.

[26] W. Wolberg., W. Street, and O. Mangasarian, "Machine learning techniques to diagnose breast cancer from image-processed nuclear features of fine needle aspirates." *Cancer Lett*, vol. 77, no. 2-3, pp. 163–71, 1994. [Online]. Available: http://www.biomedsearch.com/nih/Machine-learning-techniques-to-diagnose/8168063.html

[27] K. A. Andersen and J. N. Hooker, "Bayesian logic," *Decision Support Systems*, vol. 11, no. 2, pp. 191–210, Feb 1994.

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

[29] J. N. Hooker, "Mathematical programming methods for reasoning under uncertainty," in *Operations Research 1991*, ser. Operations Research Proceedings 1991, W. Gaul, A. Bachem, W. Habenicht, W. Runge, and W. Stahl, Eds., vol. 1991. Springer Berlin Heidelberg, 1992, pp. 23–34.

[30] Z. Ognjanovi, A. Perovi, and M. R. kovi, "An axiomatizatin of qualitative probability," *Acta Polytechnica Hungarica*, vol. 1, no. 5, pp. 105–110, 2008.

[31] P. Hansen, B. Jaumard, M. P. de Aragão, F. Chauny, and S. Perron, "Probabilistic satisfiability with imprecise probabilities," *International Journal of Approximate Reasoning*, vol. 24, no. 23, pp. 171 – 189, 2000.

[32] P.S.S.Andrade, J. da Rocha, D.P.Couto, A.C.Teves, and F.G.Cozman, "A toolset for propositional probabilistic logic," in *Encontro Nacional de Inteligłncia Artificial*. SBC, 2007, pp. 1371–1380.

[33] K. McGarry, "A survey of interestingness measures for knowledge discovery," *Knowl. Eng. Rev.*, vol. 20, no. 1, pp. 39–61, 2005.

[34] T. D. Bie, "Maximum entropy models and subjective interestingness: an application to tiles in binary databases." *Data Mining and Knowledge Discovery*, vol. 23, no. 3, pp. 407–446, 2011.

[35] G. Dong and J. Li, "Interestingness of discovered association rules in terms of neighborhood-based unexpectedness," in *Research and Development in Knowledge Discovery and Data Mining*, ser. Lecture Notes in Computer Science, X. Wu, R. Kotagiri, and K. Korb, Eds. Springer Berlin Heidelberg, 1998, vol. 1394, pp. 72–86.

[36] A. Silberschatz and A. Tuzhilin, "What makes patterns interesting in knowledge discovery systems," *IEEE Transactions on Knowledge and Data Engineering*, vol. 8, no. 6, pp. 970–974, 1996.

[37] S. Jaroszewicz, T. Scheffer, and D. A. Simovici, "Scalable pattern mining with bayesian networks as background knowledge." *Data Min. Knowl. Discov.*, vol. 18, no. 1, pp. 56–100, 2009.

[38] J. Pearl, *Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference*. San Francisco: Morgan Kaufmann, 1988.

[39] F. Cozman, "Graphical models for imprecise probabilities," *Int. J. Approx. Reasoning*, vol. 39, no. 2-3, pp. 167–184, 2005.

[40] B. Tessem, "Interval probability propagation," *International Journal of Approximate Reasoning*, no. 7, pp. 95–120, 1992.

[41] M. P. Wellman, "Fundamental concepts in qualitative probabilistic networks," *Artificial Intelligence*, vol. 44, no. 3, pp. 257–303, 1990.

[42] B. V. Balaji and V. V. Rao, "Improved classification based association rule mining," *Intl. Journal of Advanced Research in Computer and Communication Engineering*, vol. 2, no. 5, pp. 2211–2221, May 2013.

[43] F. G. Cozman, C. de Campos, and J. da Rocha, "Probabilistic logic with strong independence," in *Advances in Artificial Intelligence, IBERAMIA-SBIA 2006, Brazilian Symposium on Artificial Intelligence*, ser. Lecture Notes in Computer Science, J. Sichman, H. Coelho, and S. Rezende, Eds., vol. 4140, 2006, pp. 612–621.

[44] J. Desrosiers and M. Lubbecke, *Column Generation*. Boston, USA: Springer, 2005, ch. A Primer in Column Generation, pp. 1–32.

[45] L. C. v. d. Gaag, "Computing probability intervals under independency constraints," in *Proceedings of the Sixth Annual Conference on Uncertainty in Artificial Intelligence*. New York, NY, USA: Elsevier Science Inc., 1991, pp. 457–466.