Expert Systems with Applications 36 (2009) 5137–5145

Contents lists available at ScienceDirect

Expert Systems with Applications

journal homepage: www.elsevier.com/locate/eswa

Interestingness filtering engine: Mining Bayesian networks for interesting patterns

Rana Malhas [a], Zaher Al Aghbari [b,*]

[a] Department of Computer Science & Engineering, University of Qatar, Doha, Qatar
[b] Department of Computer Science, University of Sharjah, P.O. Box 27272, Sharjah, United Arab Emirates

ARTICLE INFO

Keywords:
Association rules
Interestingness
Bayesian networks
Data mining

ABSTRACT

In this paper, we present a new measure of interestingness to discover interesting patterns based on the user's background knowledge, represented by a Bayesian network. The new measure (sensitivity measure) captures the sensitivity of the Bayesian network to the patterns discovered by assessing the *uncertainty-increasing potential* of a pattern on the beliefs of the Bayesian network. Patterns that attain the highest sensitivity scores are deemed interesting. In our approach, mutual information (from information theory) came in handy as a measure of uncertainty. The Sensitivity of a pattern is computed by summing up the mutual information increases incurred by a pattern when entered as evidence/findings to the Bayesian network. We demonstrate the strength of our approach experimentally using the KSL dataset of Danish 70 year olds as a case study. The results were verified by consulting two doctors (internists).

1. Introduction

A major problem faced by all association rule mining algorithms is their production of a large number of rules which incurred a secondary mining problem; namely, mining interesting association rules. The problem is compounded by the fact that 'common knowledge' discovered rules are not interesting, but they are usually strong rules with high support and confidence levels – the classical measures in Agrawal, Imielinski, and Swami (1993).

The main objective of this paper is to develop an *Interestingness Filtering Engine* (IFE) that leverages background knowledge, represented by a Bayesian network, to discover interesting patterns in datasets. A pattern is considered interesting if it is *unexpected* or *surprising* to the user (Silberschatz & Tuzhilin, 1995, 1996). A new *interestingness measure* is defined to capture the *sensitivity* of the Bayesian network (beliefs) to the patterns discovered. Patterns that attain the highest sensitivity scores are deemed interesting. For this reason, the new Interestingness Measure is called *Sensitivity*.

The IFE utilizes Bayesian networks in two perspectives: The first views the net as a causality/dependence representation of the joint probability distribution of all attributes involved in the user's preliminary set of beliefs (background knowledge). The second perspective views the Bayesian network as a probabilistic inference engine that can infer the global effect of a frequent itemset on the belief network with the aid of the *Sensitivity measure*.

The Sensitivity measure should measure the uncertainty-increasing potential of a pattern; that is the extent to which a pattern alters the beliefs of the Bayesian network to more uncertain (unexpected) probabilities, when that pattern is entered as new evidence/finding to the Bayesian network. Mutual Information from information theory came in handy as a measure of uncertainty or unexpectedness. Leveraging the symmetry property inherent in Mutual Information (i.e. $I(X;Y) = I(Y;X)$, where $I(X;Y)$ is the Mutual Information between two random variables X and Y), *the Sensitivity (interestingness) of a pattern is computed by summing up the mutual information increases incurred by a pattern when entered as finding(s) to the Bayesian network.* Only increases in mutual information are considered because we are in pursuit of patterns that increase the degree of variation (i.e. unexpectedness/uncertainty) in the posterior probabilities (beliefs) of the nodes in the Bayesian network.

A case study – the KSL dataset of Danish 70 year olds – was used to analyze and verify the experimental results obtained when applying the IFE and its sensitivity measure, which exhibited a strong capability in discovering *interesting (unexpected)* patterns that are not 'common knowledge' patterns.

2. Related work and motivation

Since the inception of the classical Apriori algorithm (Agrawal et al., 1993) for mining association rules, development of interestingness measures has been a vigilant area of research. Some approaches used *objective measures* of interestingness, while others used *subjective measures*. Both interestingness measures have been used to either prune non-interesting rules (after producing all

* Corresponding author.
E-mail addresses: rmalhas@sharjah.ac.ae (R. Malhas), zaher@sharjah.ac.ae (Z.A. Aghbari).

0957-4174/$ - see front matter © 2008 Elsevier Ltd. All rights reserved.
doi:10.1016/j.eswa.2008.06.028

rules) or to discover only interesting ones (Cristofer & Simovici, 2002).

Objective measures are those that depend on the structure of the rule (pattern) discovered and the underlying data used in the discovery process (Silberschatz & Tuzhilin, 1996); such measures include: support, confidence, correlation, chi square, lift, gain ... etc. *Subjective measures*, on the other hand, rely mainly on the user who examines the patterns, where his/her prior knowledge/beliefs can adversely affect the discovery process. Most approaches using subjective measures also used some objective measures-mainly the support measure.

According to Silberschatz and Tuzhilin (1995, 1996), subjective measures were classified into two categories; namely, *unexpectedness* and *actionability*. A pattern/rule is *unexpected*, if it is 'surprising' to the user; and it is *actionable* if the user can act on it to his/her advantage.

Approaches using objective measures are not within the intended scope of this paper. Nevertheless, the reader can revert to a relatively comprehensive survey (Hilderman & Hamilton, 1999) that covers both objective and some subjective measures. As the focus of this paper is on discovering interesting patterns based on background knowledge, only related work within this context is discussed.

Three main approaches that use background knowledge in their discovery schemes were identified. The first approach is *syntax-based* as in Chen, Hsu, and Liu (1997), Chen, Hsu, Liu, and Ma (2000), Klemettinen, Mannila, Ronkainen, and Verkamo (1994), Sahar (1999); the second is *logic-based* as in Padmanabhan and Tuzhilin (1998, 2000); and the third is *probability-based* as in Jaroszewicz and Simovici (2004), Silberschatz and Tuzhilin (1995, 1996), Jaroszewicz and Scheffer (2005); noting that in Silberschatz and Tuzhilin (1995, 1996), the Bayesian probabilistic approach was one among other proposed approaches, such as the Dempster–Shafer approach and the "frequentist" probabilistic approach.

The *syntax-based approach* necessitates defining some kind of language for knowledge representation governed by a set of syntax rules, so that pair wise comparison of rules can be conducted in the discovery process; one from the set of knowledge rules and the other from the data rules. If a syntax difference is captured by the comparison – i.e. a similar rule body but a dissimilar rule head or vice versa – a data rule is considered unexpected, and hence interesting.

The *logic-based approach* is similar to the syntax-based approach since it also adopts pair wise comparison of rules; but with the difference that the comparison process looks for logical contradictions and not syntax differences between prior knowledge rules and data rules.

As for the *probability-based approach*, we think that it did not get its fair share in the pattern mining literature. A probability-based approach, basically, uses a belief system for background knowledge representation, to be able to introduce uncertainty through assigning some *degree* or *confidence factor* to each belief. Although in Silberschatz and Tuzhilin (1995, 1996) laid the ground for using a Bayesian or a Dempster–Shafer approach (reasoning under *uncertainty)*, their algorithms in Padmanabhan and Tuzhilin (1998, 2000) leveraged logical reasoning based on the user's *precise (certain)* knowledge! Two recent papers; the first by Jaroszewicz and Simovici, 2004, and the second by Jaroszewicz and Scheffer (2005) were the first to pick what Silberschatz and Tuzhilin have seeded in Padmanabhan and Tuzhilin (1998, 2000) by adopting a discovery scheme that used a Bayesian network to represent background knowledge. But they differed in the definition of their interestingness measures.

In Silberschatz and Tuzhilin (1995, 1996), the proposed definition of *interestingness* should measure *how much a pattern affects the degrees of the beliefs in a belief system*; i.e. the more a pattern disagrees with the belief system the more unexpected and hence the more interesting it is. No formal algorithm was proposed in either paper. Alternatively, in Jaroszewicz and Simovici (2004), Jaroszewicz and Scheffer (2005) the measure of *interestingness of an itemset was defined* as *the absolute difference between its support estimated from the dataset and the Bayesian network*. Itemsets with strongly diverging supports are considered interesting. We have implemented the above two inspiring definitions of interestingness (Jaroszewicz & Simovici, 2004; Silberschatz & Tuzhilin, 1996) and tested them using the KSL dataset of Danish 70 year olds.

The interestingness measure in Jaroszewicz and Simovici (2004), Jaroszewicz and Scheffer (2005) did discover relatively interesting itemsets. But, a drawback of this measure was that it only captured the partial effect of an itemset on the Bayesian network, because the measure is estimating the joint probability (i.e. support) of the itemset itself, and not the posterior probabilities conditioned on this itemset when entered to the network as new evidence (findings).

Moreover, the measure of interestingness in Silberschatz and Tuzhilin (1996) could not discover unexpected (interesting) patterns. The most interesting pattern discovered when applying this measure on the KSL dataset was *expected* in the sense that it was suggesting an association between two attributes that were assumed to be dependent in the Bayesian network.

Nevertheless, both papers motivated the work in this paper and shaped the rationale behind our newly introduced *Sensitivity* measure.

3. Definitions and notation

Using database notation:

- Let $A_1, A_2, A_3, \ldots, A_i$ be the attributes of a dataset. The domain of an attribute A_k is donated by $Dom(A_k)$. Only categorical and discrete attributes with finite domains are considered.
- Let $I, J, K, \ldots$ (uppercase letters) be the *attribute sets*, where an attribute set $I = \{A_1, A_2, A_3, \ldots, A_k\}$. The domain of an attribute set I is:

$$Dom(I) = Dom(A_1) \times Dom(A_2) \times \ldots \times Dom(A_k).$$

- Let $i, j, \ldots$ (lowercase letters) be the values from the domains of attributes and attribute sets, where $i \in Dom(I)$ is a value from the domain of the attribute set I.
- P_I denote the joint probability distribution of the attribute set I, where $P_I(i) = Pr(I = i)$ is the probability that $I = i$. Note that $\sum_{i \in Dom(I)} P_I = 1$.
- Let the pair (I,i) be an *itemset*, where I is an attribute set and $i \in Dom(I)$.
- The *support of an itemset* (I,i) in a dataset is defined as

$$\mathrm{supp}_{Data}(I, i) = P_I(i). \tag{3.1}$$

An itemset (I,i) is *frequent* if its support is greater or equal to some user-specified minimum support.

- Let BN be a *Bayesian network* over a set of attributes $H = A_1, A_2, \ldots, A_n$. The Bayesian network is a directed acyclic graph $BN = (V,E)$ with the set of vertices $V = V_{A1}, V_{A2}, \ldots, V_{An}$, and a set of edges $E \subset V \times V$. Each vertix V_A has a conditional probability distribution $P_{A_i|par_i}$, where $par_i = \{A_j : (V_{Aj}, V_{Ai}) \in E\}$ is the set of attributes corresponding to the parents of V_{Ai}.
- A Bayesian network BN over H encodes a *joint probability distribution* of H represented by

$$P_H^{BN} = \prod_{i=1}^{n} P_{A_i|par_i} \tag{3.2}$$