Information Sciences 181 (2011) 419–434

Contents lists available at ScienceDirect

Information Sciences

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

Using ontologies to facilitate post-processing of association rules by domain experts

Gunjan Mansingh [a,*], Kweku-Muata Osei-Bryson [b], Han Reichgelt [c]

[a] Department of Computing, The University of the West Indies, Mona Campus, Kingston, Jamaica
[b] School of Business, Virginia Commonwealth University, VA, USA
[c] Southern Polytechnic State University, Marietta, GA, USA

ARTICLE INFO

Article history:
Received 20 August 2008
Received in revised form 3 September 2010
Accepted 12 September 2010

Keywords:
Association rule induction
Data mining process
Healthcare
Ontologies
Objective measures
Subjective measures
Interestingness

ABSTRACT

Data mining is used to discover hidden patterns or structures in large databases. Association rule induction extracts frequently occurring patterns in the form of association rules. However, this technique has a drawback as it typically generates a large number of association rules. Several methods have been proposed to prune the set of extracted rules in order to present only those which are of interest to the domain experts. Some of these methods involve subjective analysis based on prior domain knowledge, while others can be considered to involve objective, data-driven analysis based on numerical measures that provide a partial description of the interestingness of the extracted association rules. Recently it has been proposed that ontologies could be used to guide the data mining process. In this paper, we propose a hybrid pruning method that involve the use of objective analysis and subjective analysis, with the latter involving the use of an ontology. We demonstrate the applicability of this hybrid method using a medical database.

1. Introduction

Large databases often contain interesting, hidden patterns of knowledge that can be uncovered using data mining techniques [7,33,47]. Often the number of patterns that are discovered can be quite large. This is particularly true for association rule (AR) induction [52], where the large number of extracted association rules makes it difficult for decision makers to process, interpret and utilize them in the decision making process. To reduce the effort required to identify interesting rules, researchers have offered approaches that aim to minimize the number of rules generated [7,31,44]. A review of the literature shows that both subjective and objective methods are essential in examining the interestingness of rules and for pruning the extracted association rules [7,17]. In subjective analysis, prior domain knowledge is used to determine unexpectedness in the extracted association rules [2,22,27,31,44]. In objective analysis, the numerical measures (e.g. *Support*, *Confidence*) associated with each AR is used to prune and rank the ARs [2,45]. Researchers agree that both methods for determining interestingness are important, however the methods for integrating them are few [7,17].

The interestingness of an extracted AR can be described in terms of its unexpectedness and actionability. Expected rules confirm prior domain knowledge and are essentially *known*; unexpected rules are *novel* rules (i.e. those which were previously *unknown*) which may contradict the user's existing knowledge. A rule is actionable if a domain expert can use it to their advantage [31]. In this study, our focus is on examining the interestingness of rules based on unexpectedness. Since prior domain knowledge is required to determine unexpectedness in terms of which rules are *known* or *novel*, our approach

[*] Corresponding author. Tel./fax: +1 (876) 702 4455.
E-mail address: gunjan.mansingh@uwimona.edu.jm (G. Mansingh).

0020-0255/$ - see front matter © 2010 Elsevier Inc. All rights reserved.
doi:10.1016/j.ins.2010.09.027

involves the use of ontologies which can effectively represent this domain knowledge [20]. It should also be noted that while the use of ontologies have been previously proposed for the knowledge discovery and data mining (KDDM) process [3,10,29,34,41,43,51], in this study we focus on their use in facilitating the processing of ARs in a manner that significantly reduces the cognitive burden on the domain expert(s).

The paper is organized as follows. In Section 2, we review the background literature relevant to this research. Section 3 describes a method of using an ontology to partition a set of generated association rules into meaningful partitions and integrating the newly extracted knowledge into the ontology. Section 4 demonstrates the applicability of the method to a medical database. Section 5 provides concluding remarks and directions for future research.

2. Literature review

2.1. Association rule induction

An association rule (AR) shows relationships among items in a transaction of a database [11]. These patterns or rules have been used for various purposes, for example, to enrich static schemas such as Entity Relationship models, and to improve marketing campaigns by grouping products that target particular market segments [16,49]. The AR induction process can be divided into two steps. First the database is scanned to extract all the itemsets that satisfy a user-specified minimum support criterion (see Table 1). Then each AR that describes an association between items in a transaction that occur frequently is extracted based on a user-specified minimum *confidence* criterion [1,39]. The process usually results in a large number of ARs. In order to reduce the number of rules generated, the minimum *confidence* threshold value can be increased. However, setting the minimum *confidence* level too high may prevent the identification of some important ARs.

The Apriori algorithm is one of most commonly used methods for AR induction [1]. Although Apriori has been successfully applied in many cases, it does have performance problems, and so several algorithms have been proposed to improve Apriori's performance [4,5,42,46]. In both the Apriori algorithm and in subsequent improvements, the minimum *support* count and the minimum *confidence* value have to be supplied by the data mining expert. Settings for appropriate *support* and *confidence* values is often a matter of trial and error [50]. If threshold values are set too high, only a small number of results will be generated; if they are set too low, too many results will be generated, thus wasting computational time.

2.1.1. Subjective analysis of interestingness

Subjective analysis requires user involvement in finding interesting ARs [22,27] based on their unexpectedness and actionability [31,44]. Both Liu et al. [31] and Silberschatz and Tuzhilin [44] proposed methods to assess unexpectedness that requires the user to specify domain knowledge, which is then used to determine whether a discovered rule is unexpected. Unexpected rules are presented to the user who ultimately decides which are retained and which are discarded. To specify the domain knowledge, Liu et al. [31] created a simple specification language that involves three types of prior knowledge: *general impressions*, *reasonably precise concepts* and *precise knowledge*. *General impressions* involve the user's vague beliefs about the associations between certain concepts; *reasonably precise concepts* specify both the association between concepts and the direction of the association; and *precise knowledge* includes the *support* and the *confidence* values. Liu et al. [31] use *general impression* and *reasonably precise concept* knowledge to categorize the extracted rules as conforming, unexpected consequent, unexpected condition and both sides (i.e. antecedent and consequent) unexpected rules.

2.1.2. Objective analysis of interestingness

Several objective measures to determine the interestingness of an association rule exist in the literature [2,8,35,45] (see Table 1). According to Tan et al. [45], data mining practitioners tend to apply objective measures without considering alternatives, which leads to difficulties in interpreting rules because objective measures may provide conflicting information about the interestingness of a rule.

Table 1
Objective measures.

Measures	Formula	Description	Symmetric
Support (S)	$P(X \cap Y)$	Probability of cases that contain both X and Y	Yes
Confidence (C)	$\frac{P(X \cap Y)}{P(X)}$	Confidence for a rule is defined in terms of the percentage of cases that contain X and Y given that X is being observed	No
Lift or interest (I)	$\frac{P(X \cap Y)}{P(X)P(Y)}$	Lift is a ratio between the observed support of X and Y and its expected support under the assumption that X and Y are independent	Yes
Conviction (V)	$\frac{P(X)P(\overline{Y})}{P(X \cap \overline{Y})}$	Conviction measures independence of X and Y	Yes
Piatetsky–Shapiro's (PS)	$P(X \cap Y) - P(X)P(Y)$	Piatetsky–Shapiro's is an interestingness measure	Yes
Reliability (r)	$\left\lvert \frac{P(X \cap Y)}{P(X)} - P(Y) \right\rvert$	Reliability measures independence of X and Y	No