Interesting patterns extraction using prior knowledge

Laurent Brisson

Interesting Patterns Extraction Using Prior Knowledge

Laurent Brisson

Laboratoire I3S - Université de Nice, 06903 Sophia-Antipolis, France
brisson@i3s.unice.fr

Abstract. One important challenge in data mining is to extract interesting knowledge and useful information for expert users. Since data mining algorithms extracts a huge quantity of patterns it is therefore necessary to filter out those patterns using various measures. This paper presents IMAK, a part-way interestingness measure between objective and subjective measure, which evaluates patterns considering expert knowledge. Our main contribution is to improve interesting patterns extraction using relationships defined into an ontology.

1 Introduction

In most data mining projects, prior knowledge is implicit or is not organized as a structured conceptual system. We use ExCIS framework [1] which is dedicated to data mining situations where the expert knowledge is crucial for the interpretation of mined patterns. In this framework the extraction process makes use of a well-formed conceptual information system (CIS) for improving the quality of mined knowledge. A CIS is defined by Stumme [8] as a relationnal database together with conceptual hierarchies.

Numerous works focused on indexes that measure the interestingness of a mined pattern [3]. They generally distinguished objective and subjective interest. Silberschatz and Tuzhilin [6] proposed a method to define unexpectedness and actionability via belief systems while Liu [3] developed a method that use user expectations. In this paper we present an interestingness measure called IMAK, which evaluates extracted patterns according prior knowledge. The novelty of this approach lies in the use of a Conceptual Information System in order to extract rules easily comparable with knowledge. This ontology based approach for unexpected and actionable patterns extraction differs from works on interestingness measures.

The paper is organized as follows. In section 2, we study related works. Section 3 focus on interesting patterns extraction. Section 4 concludes the paper.

2 Interestingness Measures

Among all indexes that measure the interestingness of a mined pattern there are measures of objective interestingness such as confidence, coverage, lift, success

rate while unexpectedness and actionability are proposed for subjective criteria. In this section we presents only subjective interestingness measures.

2.1 What Makes Patterns Interesting?

Silberschatz [7] presents a classification of measures of interestingness and identifies two major reasons why a pattern is interesting from the subjective (user-oriented) point of view:

- Unexpectedness: a pattern is interesting if it is surprising to the user
- Actionnability: a pattern is interesting if the user can do something with it to his or her advantage

Therefore a pattern can be said to be interesting if it is both unexpected and actionable. This is clearly a highly subjective view of the patterns as actionability is dependent not only on the problem domain but also on the user's objectives at a given point in time [4]. According to the actionability criteria, a model is interesting if the user can start some action depending on it [7]. On the other hand, unexpected models are considered interesting since they contradict user expectations which depend on his beliefs.

2.2 Belief System [6]

Silberschatz and Tuzhilin proposed a method to define unexpectedness via belief systems. In this approach, there are two kinds of beliefs: soft beliefs that the user is willing to change if new patterns are discovered and hard beliefs which are constraints that cannot be changed with new discovered knowledge. Consequently this approach assumes that we can believe in certain statements only partially and some degree or confidence factor is assigned to each belief. A pattern is said to be interesting relatively to some belief system if it "affects" this system, and the more it "affects" it, the more interesting it is.

2.3 User Expectations [3]

User expectations is a method developed by Liu. User had to specify a set of patterns according to his previous knowledge and intuitive feelings. Patterns had to be expressed in the same way that mined patterns. Then Liu defined a fuzzy algorithm which matches these patterns. In order to find actionable patterns, the user has to specify all actions that he can take. Then, for each action he specifies the situation under which he is likely to run the action. Finally, the system matches each discovered pattern against the patterns specified by the user using a fuzzy matching technique.

3 Interesting Patterns Extraction

3.1 Knowledge Properties

We chose to express knowledge like "if ... then ..." rules in order to simplify comparison with extracted association rules. Each knowledge has some essential properties to select the most interesting association rules:

- Confidence level: 5 different values are available to describe knowledge confidence according a domain expert. These values are range of confidence value: 0-20%, 20-40%, 40-60%, 60-80% and 80-100%. We call confidence the probability the consequence of a knowledge occurs when the condition holds.
- Certainty:
  - Triviality: cannot be contradicted
  - Standard knowledge: domain knowledge usually true
  - Hypothesis: knowledge the user want to check

Since our project deals with data from the "family" branch of the French national health care system (CAF), our examples are related to CAF domain. Let's consider the following knowledge:

KNOWLEDGE 1
Objective='To be paid' $\wedge$ Allowance='Housing Allowance' $\wedge$ Distance='0km' $\rightarrow$ Contact='At the agency'

- Confidence level: 60-80%
- Certainty: Hypothesis

3.2 IMAK : An Interestingness Measure According Knowledge

We propose an interestingness measure *IMAK* which considers actionnalibity (using certainty knowledge property) and unexpectedness (using generalization relationships between ontology concepts). Although unexpected patterns are interesting it's necessary to consider actionable expected patterns. In our approach we deal with actionnability using knowledge certainty property:

- If a pattern match a trivial knowledge it isn't actionable since actions concerning trivial knowledge are most likely known
- Since user knowledge define his main points of interest, a pattern matching standard knowledge could be actionable
- If a pattern match a hypothesis, it is highly actionable

IMAK only considers confidence as objective interestingness measure. Consequently it can't be applied on rules with lift ≤ 1. It makes no sense to compare with knowledge rules whose antecedent and consequent aren't positively correlated.

Our measure describes four levels of interest:

- none: uninteresting information
- low: confirmation of standard knowledge
- medium: new information about a standard knowledge / confirmation of a hypothesis
- high: new information about a hypothesis

As you can see in table 1, IMAK value increases when a pattern matches a hypothesis and decreases when it matches a triviality. Furthermore IMAK value increases when a pattern is more general than a knowledge or when its confidence

Table 1. IMAK values

Knowledge Certainty / Pattern is	Triviality	Standard knowledge	Hypothesis
Case 1. Pattern with better confidence level than knowledge			
more general	medium	high	high
similar	none	low	medium
more specific	none	medium	high
Case 2. Pattern and knowledge with similar confidence level			
more general	low	medium	high
similar	none	low	medium
more specific	none	low	medium
Case 3. Pattern with lesser confidence level than knowledge			
more general	none	none	low
similar	none	low	medium
more specific	none	none	low

level is the best. Generalization level of a rule compared to a knowledge is defined with the help of the embedded ontology in ExCIS framework.

3.3 Experimental Results

Let's consider the knowledge rule 1, and the two following extracted rules:

Extracted rule 1
Objective='To be paid' ∧ Allowance='Housing Allowance' → Contact='At the agency' [confidence=20%]

Extracted rule 2
Objective='To be paid' ∧ Allowance='Housing Allowance'
∧ Distance='LessThan30km' → Contact='At the agency' [confidence=95%]

Rule 1 is a generalization of the knowledge. But its confidence is lesser than knowledge confidence level. Consequently *IMAK* value is "low" since the knowledge is a "hypothesis" (see table 1 column 3 line 7).

Rule 2 is also a generalisation of the knowledge. Its confidence is better than than knowledge confidence level. Consequently *IMAK* value is "high" since the knowledge is a "hypothesis" (see table 1 column 3 line 1). Now let's consider the rule:

Extracted rule 3
Objective='To be paid' ∧ Allowance='Student Housing Allowance'
∧ Distance='0km' → Contact='At the agency' [confidence=75%]

Rule 3 is more specific than knowledge and its confidence is similar. Consequently *IMAK* value is "medium" since the knowledge is a "hypothesis" (ref table 1 column 3 line 6).

We apply IMAK measure on 5000 rules extracted by several runs of CLOSE algorithm [5]. CAF experts couldn't deal with such a number of rules. However after having defined their knowledge we could present them a hundred of interesting rules classified into few categories.

4 Conclusion

We presented IMAK, an interestingness measure, which evaluates extracted patterns according to prior knowledge. Some works on subjective interestingness measures [2,3,6] use templates or beliefs in order to express knowledge. Our contribution is to improve interesting patterns extraction using relationships defined into an ontology [1]. IMAK measure doesn't make syntaxic matching but uses semantic relationships between concepts, analyzes rules cover, compares confidence level and takes into account the knowledge certainty. Consequently it is part-way between objective and subjective measure. In future works we plan to compute IMAK using ontology relationships which aren't generalization/specialization relationships and to evaluate our measure on a less subjective application domain.

References

1. L. Brisson, M. Collard and N. Pasquier. *Improving the Knowledge Discovery Process Using Ontologies*. Proceedings of Mining Complex Data workshop in ICDM Conference, November 2005
2. M. Klemettinen, H. Mannila, P. Ronkainen, H. Toivonen and A. Verkamo. Finding interesting rules from large sets of discovered association rules In CIKM-94, 401 – 407, November 1994.
3. B. Liu, W. Hsu, L.-F. Mun and H.-Y. Lee. *Finding Interesting Patterns using User Expectations*. Knowledge and Data Engineering, 11(6):817-832, 1999.
4. K. Mcgarry *A Survey of Interestingness Measures for Knowledge Discovery* The knowledge engineering review, vol. 00:0, 1-24, 2005.
5. Pasquier N., Taouil R., Bastide Y., Stumme G. and Lakhal L. *Generating a Condensed Representation for Association Rules.* Journal of Intelligent Information Systems, Kerschberg L., Ras Z. and Zemankova M. editors, Kluwer Academic Publishers
6. A. Silberschatz and A. Tuzhilin. *On Subjective Measures of Interestingness in Knowledge Discovery.* Proceedings 1st KDD conference, pp. 275-281, august 1995.
7. A. Silberschatz and A. Tuzhilin. *What Makes Patterns Interesting in Knowledge Discovery Systems.* IEEE Transaction On Knowledge And Data Engineering, 8(6):970-974, december 1996.
8. G. Stumme. *Conceptual On-Line Analytical Processing.* K. Tanaka, S. Ghandeharizadeh and Y. Kambayashi editors. Information Organization and Databases, chpt. 14, Kluwer Academic Publishers, pp. 191-203, 2000.