An Interactive Approach for the Post-processing in a KDD Process

P.A. Potes Ruiz, B. Kamsu-Foguem, B. Grabot

Laboratoire Génie de Production / INP-ENIT - Université de Toulouse
47, Avenue d'Azereix, BP 1629, F-65016 Tarbes Cedex – France

{paula.potesruiz, bernard.kamsu-foguem, bernard.grabot}@ enit.fr

Abstract. Association rule mining is a technique widely used in the field of data mining, which consists in discovering relationships and/or correlations between the attributes of a database. However, the method brings known problems among which the fact that a large number of association rules may be extracted, not all of them being relevant or interesting for the domain expert. In that context, we propose a practical, interactive and helpful guided approach to visualize, evaluate and compare the extracted rules following a step by step methodology, taking into account the interaction between the domain expert and the data mining expert.

Keywords: Knowledge Discovery from Databases, Association Rules Mining, Post processing phase, Interactivity, Decision Support System.

1 Introduction

Advances in information and storage technology have promoted the interest of companies for research works like knowledge discovery from databases. Particularly, the generalisation of the ERP (Enterprise Resource Planning) in industrial environments, make available a large amount of information. Hence, data mining techniques can be used to process this information and extract new knowledge, potentially useful to support decision-making. Nevertheless, this extraction should include a post-processing phase assessing the usefulness and reliability of the results, before their validation [**Erreur ! Source du renvoi introuvable.**]. We propose in this paper an interactive approach for this post-processing phase, controlled by a domain expert and a data-mining (DM) expert.

2 Knowledge Discovery from Databases (KDD)

The knowledge extraction approaches have developed new intelligent tools, more efficient than traditional data analysis methods for discovering new knowledge in an industrial context. Knowledge Discovery from Databases KDD is defined as a "*non-*

trivial process of identifying valid, novel, potentially useful and ultimately understandable patterns in data" [**Erreur ! Source du renvoi introuvable.**], in order to create a significant competitive advantage in companies. Given the great potential of the available data as a source of new knowledge [**Erreur ! Source du renvoi introuvable.**], KDD has become essential in many industrial fields, including product and process design, materials planning, quality control, scheduling, maintenance, customer relationship management, etc.

The general process involves three main phases: pre-processing, data mining, and post-processing.

Pre-processing phase: this phase requires a special attention in order to have reliable data before applying the extraction algorithms, guaranteeing therefore the quality of the results generated. Data cleaning, data discretization, data reduction or data transformation techniques can be used in that purpose.

Data mining phase: Data mining consists in applying data analysis and discovery algorithms to find hidden knowledge (relations or patterns) in large volumes of information [**Erreur ! Source du renvoi introuvable.**, **Erreur ! Source du renvoi introuvable.**]. Our focus is on the association rules mining approach [**Erreur ! Source du renvoi introuvable.**] to discover relationships between a set of attributes (or items) in a database. The obtained relationships are based on the co-occurrence of attributes [**Erreur ! Source du renvoi introuvable.**] showing correlation, but not a cause.

An association rule is formally defined as a relationship between two itemsets through relations of the form "If X, then Y", denoted as $(X \rightarrow Y)$, where $X, Y \in I$ and $X \cap Y = \emptyset$. X is usually called hypothesis and Y conclusion, i.e. the presence of X allows to conclude on the presence Y. Two classical measures are usually related to assess the discovered association rules: support and confidence. The support of a rule is the proportion of transactions in a database that contain both X and Y, and the confidence indicates the proportion of transactions containing Y among those containing X.

$$Support(X \rightarrow Y) = \mathrm{P}(X \cap Y) \tag{1}$$

$$Confidence(X \rightarrow Y) = P(Y|X) = \frac{Suport(X \rightarrow Y)}{Suport(X)} \tag{2}$$

Post-processing phase: the last phase of the process is the analysis and interpretation of discovered information. Over the years, many efforts have focused on improving algorithmic performance (in terms of execution time and memory consumption) but this phase has been surprisingly neglected. The post-processing of the results is nevertheless becoming increasingly important in companies, in order to find and validate the most interesting rules for each specific problem.

We present in more details in the next sections an original approach aiming at an easier interpretation and comparison of the obtained rules, their interest being decided with the assistance of a domain expert to ensure the relevance of the extraction process in a given company.

3 An Interactive Post-processing phase in the KDD

Four notions characterize the interest of extracted models [**Erreur ! Source du renvoi introuvable.**]: validity, novelty, usefulness and comprehension by the user. The models should validate the analysed data set and to some extent, new data sets; bring new knowledge to the user; be useful to support decision making, and be understandable by the decision maker. We focus especially here on the usefulness and comprehension by the user, within an interactive approach, underlining the indispensable role of the human in the process [**Erreur ! Source du renvoi introuvable.**].

The phases of a user-centered KDD process are shown in Fig. 1. The post-processing phase, which is in our opinion of specific interest, is necessary to evaluate and filter extracted rules, and we consider that it should not be automated.

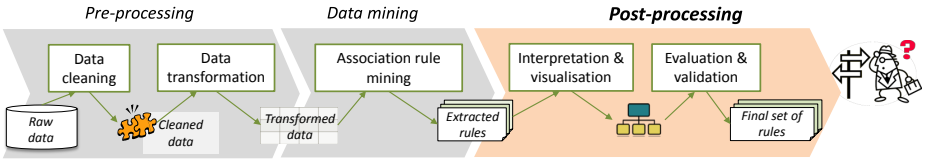

Fig. 1. The post processing phase in the KDD process.

3.1 Interaction between the domain expert and DM expert

In practice, it is difficult to find a DM expert who is also an expert in the industrial domain considered. We address in this section the importance of the collaboration between the experts in the process, to guarantee the quality of results and to make the knowledge extraction process more relevant for the enterprise.

The domain expert is notably the person who knows the field and is responsible for decision-making. In contrast, the DM expert develops and manages the data mining techniques that will obviously support decision. In that context, we want to involve the domain expert in the interpretation and evaluation of the results obtained by the DM expert, and then in the validation of the elements of interest of these results. Interaction in the post-processing phase is a means for sharing knowledge [**Erreur ! Source du renvoi introuvable.**]. Inspired from [**Erreur ! Source du renvoi introuvable.**], we suggest a model that articulates the knowledge between these two experts (**Fig. 2**).

The KDD cycle related to the DM expert (right path in **Fig. 2**) concerns firstly a phase of exchange between experts, to define the initial problem. The pre-treatment and DM phases are then carried out. Finally, the post-processing phase is considered to interpret and evaluate the results obtained with the assistance of a domain expert. On the other hand, the domain expert centered cycle (left path in **Fig. 2**) concerns the post-processing of results derived from the DM phase, then a validation according to the needs and/or expectation of the domain, a decision making and an integration in the industrial field for improving existing processes. Finally, a positive and/or negative feedback outcome of this cycle must be carried out to the DM expert to enhance the new DM tasks, during a new knowledge extraction cycle.

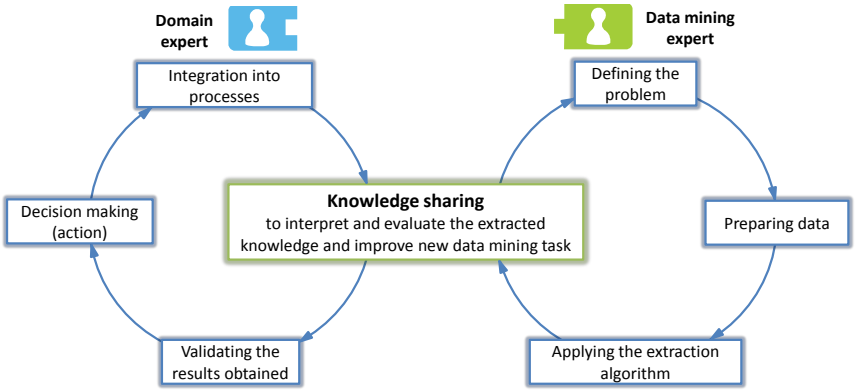

Fig. 2. Knowledge sharing between the domain expert and data-mining expert

3.2 Interpreting and evaluating extracted knowledge

We suggest three ways to evaluate the association rules, inspired from a classification presented in [**Erreur ! Source du renvoi introuvable.**]: *i)* an "objective evaluation" (based on the support and confidence), *ii)* a "semantic evaluation" (based on the domain knowledge), and *iii)* a "subjective evaluation" (based on the goals and beliefs of the domain expert).

Objective evaluation. This traditional knowledge evaluation is performed during the association rules mining. Although other statistical measures have been proposed in the literature, an objective rule evaluation is often done by determining the rules that have a support and a confidence superior or equal to user-defined thresholds. So, we focus here on the interpretation of *minsup* and *minconf* thresholds of the support and confidence of the obtained rules.$

The minsup and minconf thresholds are predefined for applying an extraction algorithm (here, the well known Apriori algorithm [**Erreur ! Source du renvoi introuvable.**]). Indeed, they are a first way to evaluate the extracted rules, without guarantee of their usefulness. Choosing the optimal levels of these parameters is a difficult task. A very low minsup would lead to a combinatorial explosion of number of candidate itemsets; on the contrary, a very high minsup would prevent the appearance of association rules containing rare but interesting attributes [**Erreur ! Source du renvoi introuvable.**]. If minsup=0, each examined transaction would be expressed by a different rule (no generalization is performed), otherwise if minsup=1, a single rule would be generated under condition that all the transactions contain the same itemset. The minconf has a different interest: it shows the validity of a rule, i.e. up to what point the conclusion part is related to the hypothesis part. A high minconf allows to generate very robust rules, but in practice, these rules are usually well known by domain experts. Instead, the rules with low confidence may seem inconsistent, but may also express unusual but interesting situations.

In practice, the efficiently treatment of attributes characterizing the transactions requires to test different thresholds since rare rules may be more interesting than frequent ones.

In that context, many studies on association rules evaluation were limited to determine the interest of a rule from a statistical point of view, leaving a lot of inconsistent rules, or just not interesting rules from the point of view of domain expert user.

We propose following an attempt to complete this classic rules evaluation, which seems us an insufficient evaluation criteria in filtering results obtained (in terms of volume and quality). As an alternative to the limits of objective measures (support-confidence), we suggest a semantic and subjective evaluation to provide new and more relevant knowledge to the domain expert user.

Semantic evaluation. It facilitates the evaluation of the interest of a rule according to the domain knowledge. Since these domain knowledge have multiple facets and may be complex, it is difficult to give generic guidelines here, except on an original point: combining objective and semantic evaluation may allow to diagnose some aspects of coherence and consistency of the database concerned. We focus here on the interpretation and visualization of results in order to draw conclusions and suggest actions to the domain experts. These aspects will indeed improve the understanding by the user, a notion that characterizes to some extent the interests of rules.

In this regard, we propose to use the following step-by-step approach (illustrated in section 4) as a methodology to interpret and understand the extracted rules: i) analysing "elementary" rules (involving only two attributes), ii) expressing each attribute analysed by a question, iii) expressing the problem addressed by each rule by combining the questions, iv) interpreting the support and confidence of rules, v) analysing the potential use of each rule for improving the industrial processes, vi) checking whether the reverse rule is, or should be, present, indeed, analysing the rules (present but also absent), given their support and confidence, allows to identify inconsistencies in the databases (i.e. typing errors, data entry errors or anomalies defining the attributes), vii) analysing more complex rules by comparison with the elementary ones through three logical operations, denoted here as *extension* (of hypothesis or conclusion part of rules), *permutation* (of attributes between hypothesis and conclusion part of rules) and *junction* (of the hypothesis or conclusion part of rules), and then using the same steps described above, viii) representing an overall structure of the extracted rules (indicating the relationship between the identified rules), thereby facilitating understanding and a visual exploration of the mined rule set by users, ix) formalising a "metarule" to generalize a rule-set and provide a new abstraction level grouping the rules. We intend to summarize the mined rule set from a general to a specific level (graphical model). Thus, rules of an upper level provide a general overview of the knowledge (i.e. elementary rules) whereas rules of a lower level are more specific.

Subjective evaluation. It is related to looking for specific types of rules according to the user expectations (domain expert). Indeed, we consider that structuring the rules facilitates a visual exploration and assist the expert in this validation step.

In our KDD process, the target knowledge is not particularly predetermined during the extraction algorithm application, unlike others techniques constraining the number of items and/or determining what items are in the hypothesis or conclusion part. However, a domain expert user in a given situation has usually an idea on the type of rule that he/she expects, mainly rules suggesting actions for decision making.

Mining algorithms like Apriori [**Erreur ! Source du renvoi introuvable.**] allow to mine different types of rules, including rules that might be expected, but others that may be completely unexpected by the user. Unexpected rules can be also considered highly interesting and advantageous, providing to the user new knowledge.

In the literature, there are different techniques to perform this subjective evaluation of extracted rules. A study of several techniques based on knowledge/user expectations have been detailed in [**Erreur ! Source du renvoi introuvable.**]; indeed there are several formalisms used to represent such knowledge and filter the rules (i.e. templates, beliefs, meta-rules, queries, taxonomies and ontologies).

The user expectations and a query/answering mechanism. A visual representation of association rules facilitates the interaction with the user and make particularly convenient the process of modelling a query (user expectation), and then the filtering process. A query Q relates to a rule skeleton, describing the structure a priori of the interesting rules for the user among the extracted rules. A query/answering mechanism will look for "response" rules to sort a final set of potentially interesting rules.

Various types of rules discovered. Let X be a set of association rules extracted and Q a user query. Regarding to the structure of the extracted rules, [**Erreur ! Source du renvoi introuvable.**] suggests to distinguish between four sets of potentially interesting rules:

- Conforming rules: an extracted rule $X_i \in X$ is conform to the user query Q if both hypothesis and conclusion parts of X_i are consistent with respect to Q.
- Unexpected conclusion rules: a discovered rule $X_i \in X$ has an unexpected conclusion with respect to Q if the hypothesis of X_i is consistent with Q, but not the conclusion part. Unexpected conclusion rules show types of rules that may be inconsistent with the existing knowledge.
- Unexpected hypothesis rules: a discovered rule $X_i \in X$ has an unexpected hypothesis with respect to Q if the conclusion of X_i is consistent with Q, but not the hypothesis part. Unexpected antecedent rules can show other hypothesis that can lead to the same result or conclusion.
- Both-side unexpected rules: a discovered rule $X_i \in X$ is both-side unexpected with respect to Q if both the hypothesis and conclusion part of the rule X_i are not consistent with Q.

1 Application Example

We consider here a real set of reports on maintenance operations performed on equipment of production processes in a large company of the aeronautical sector. An

Excel© sheet with 5955 maintenance reports from the SAP ERP Production Maintenance module is our starting point, containing several attributes (date entered, order work number, frequency, nature, priority, equipment, model, analytical section, ..).

A first discussion with the maintenance expert allowed us to better understand these attributes in the context. Then, the KDD cycle was carried out: the data preparation, the application of Apriori algorithm, and the post-processing phase considering the domain expert in the interpretation and validation of results. Such knowledge sharing together with the domain expert, by presenting him the first partial results of the KDD process, brought us more details of the data (for example, not take into account some attributes being manifestly without interest). In that context, improving a new DM cycle is one of our goals to meet the needs or expectations of the industrial system based on the expert.

For filtering the extracted rules, we have empirically chosen minsup=20% and minconf=90% in order to present some results, leading to the extraction of 38 frequent itemsets and 16 rules. Among the results obtained, we can consider the first 6 rules established by the algorithm as "elementary". Let us now analyse in more detail some rules taking account the support, confidence and the absence of reverse rules.

- *Rule 1:* Frequency=Semi-annual → Nature=Preventive sup=0.21 conf=1.0
  Question answered: link between "how often" and "what kind of intervention".
  Interpretation: The 21% of all interventions are preventive and performed every 6 months. Every intervention that have a semi-annual frequency concern a preventive intervention (conf= 1.0). However, preventive intervention have any other frequencies (since the reverse rule is absent).
- *Rule 2.* Production=0001 → Type of equipment=A380 sup=0.23 conf=0.97
  Question answered: link between "what site" and "what type of equipment".
  Interpretation: The 23% of interventions concern the type of equipment A380 on the production site 0001. The 97% of maintenance work on this production site correspond to this type of equipment, just 3% of the interventions on this site correspond to another equipment.
- *Rule 5.* Model=Booths → Production=0002 sup=0.35 conf=1.0
  Question answered: link between "what model" and "on which site".
  Interpretation: The 35% of all interventions correspond to the booths on the production site 0002. In fact, all operations on the booths are made on this site (conf = 1.0).

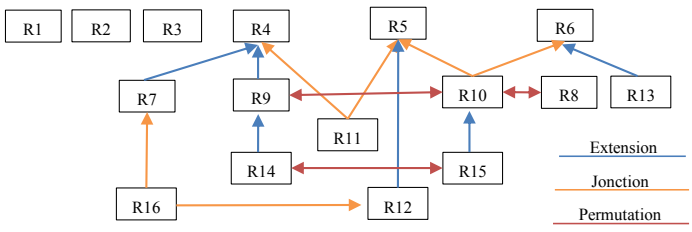

Fig. 3. Structure of final set of rules extracted

The other rules may be considered as variants of the six basic ones by mean of the extension, permutation and junction. Therefore, after that analysis, we present to the domain expert a structure of the final set of rules (**Fig. 3**); then the domain expert may make the queries on this structure in order to filter the different results, which may help to effectively guide human decision making related to processes, or simply suggest how to better structure the database. In that proposed approach, the role played by the domain expert and the quality of information available are decisive, and both affect the quality of the knowledge extracted.

2 Conclusion

The interactive approach proposed for post-processing takes into account some efforts already reported in the literature; however, its novelty contemplates the interpretation of knowledge extracted according to several factors: the support, the confidence, the presence and absence of expected rules, the reverse rules, the relationship between the rules set extracted and the frequents itemsets, and in particular the interaction between the domain expert and the DM expert. The main focus is on consideration of the domain expert in order to improve future DM process consistent with application contexts. Indeed, it is essential to understand what the user is looking in the data to be able to define the problem and apply DM techniques relevant.

References

1. R. Agrawal and R. Srikant. Fast algorithms for mining association rules in large databases. In *Proceedings of the 20th International Conference on Very Large Data Bases*, vol. 1215, pp. 487-499. Morgan Kaufmann Publishers Inc., 1994.
2. B. Baesens, S. Viaene, and J. Vanthienen. Post-processing of association rules. *DTEW Research Report 0020*, 1–18, 2000.
3. M. Ben Ayed, H. Ltifi, C. Kolski, and A M. Alimi. A user-centered approach for the design and implementation of kdd-based dss: A case study in the healthcare domain. *Decision Support Systems*, 50(1):64–78, 2010.
4. A K. Choudhary, J A. Harding, and M K. Tiwari. Data mining in manufacturing: a review based on the kind of knowledge. *Journal of Intelligent Manufacturing*, 20(5):501-521, 2009.
5. U M. Fayyad, G. Piatetsky-Shapiro, P. Smyth, and R. Uthurusamy. *Advances in knowledge discovery and data mining*. MIT Press, 1996.
6. L. Geng and H J. Hamilton. Interestingness measures for data mining: A survey. *ACM Computing Surveys*, 38(3): Article 9, 2006.
7. P. Giudici. *Applied data mining: Statistical methods for business and industry*. Wiley, 2003.
8. J A. Harding, M. Shahbaz, Srinivas, and A. Kusiak. Data mining in manufacturing: A review. *Journal of manufacturing science and engineering - transactions of the ASME*, 128(4):969-976, 2006.
9. G. Köksal, I. Batmaz, and M. Caner Testik. A review of data mining applications for quality improvement in manufacturing industry. *Expert Systems with Applications*, 38(10):13448-13467, 2011.

10. B. Liu, W. Hsu, K. Wang, and S. Chen. Visually aided exploration of interesting association rules. In N. Zhong and L. Zhou (eds), *Methodologies for Knowledge Discovery and Data Mining*, vol. 1574 of *LNCS*, pp. 380-389. Springer Berlin Heidelberg, 1999.
11. C. Marinica. *Association Rule Interactive Post-processing using Rule Schemas and Ontologies-ARIPSO*. PhD thesis, Ecole polytechnique de l'Université de Nantes, 2010.
12. H. Wang and S. Wang. A knowledge management approach to data mining process for business intelligence. *Industrial Management & Data Systems*, 108(5):622-634, 2008.