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

Paula Andrea Potes Ruiz, Bernard Kamsu-Foguem, and Bernard 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 industrial 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 (KDD). Particularly, the generalization 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 [1]. We propose in this paper an interactive approach for this post-processing phase, controlled by an industrial domain expert and a data mining 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 is defined as a "*non-trivial process of identifying valid, novel, potentially useful and ultimately understandable patterns in data*" [2], in order to create a significant competitive advantage in companies. Given the great potential of the available data as a source of new knowledge [3],

B. Grabot et al. (Eds.): APMS 2014, Part I, IFIP AICT 438, pp. 93–100, 2014.

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.

2.1 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.

2.2 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 [3, 4]. Our focus is on the association rules mining approach [5] to discover relationships between a set of attributes (or items) in a database. The obtained relationships are based on the co-occurrence of attributes [6] 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 of 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.

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

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

2.3 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 an industrial 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 [7]: 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 [8].

3.1 Interaction between the Industrial Domain Expert and a Data Mining Expert

In practice, it is difficult to find a data mining expert (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 industrial domain expert (ID 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 ID 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 [9]. Inspired by the work of Wang and Wang [9], we suggest a model that articulates the knowledge between these two experts (Fig. 1).

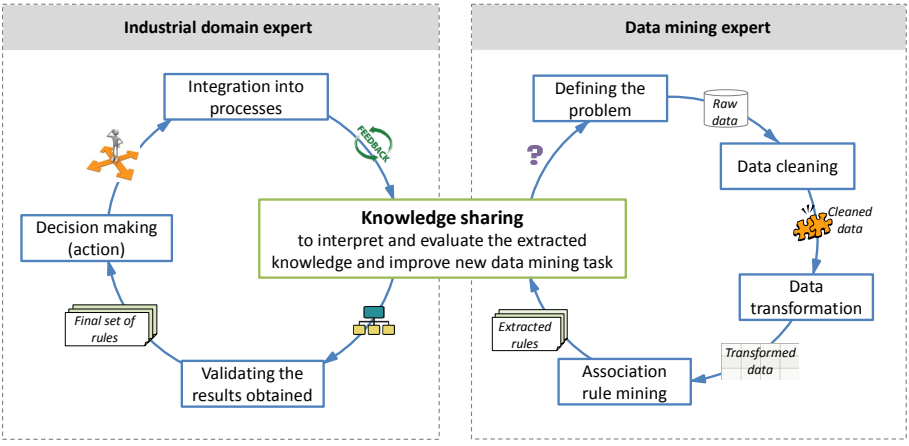

Fig. 1. Knowledge sharing between the ID expert and DM expert

The KDD cycle related to the DM expert (right path in Fig. 1) concerns firstly a phase of exchange between experts, to define the initial problem. The pre-processing phase (data cleaning and data transformation) and data mining phase (association rule mining) are then carried out. Finally, the post-processing phase is considered to

interpret and evaluate the results obtained with the assistance of an ID expert. This phase, which is in our opinion of specific interest, is necessary to filter extracted rules. We consider that it should not be automated. On the other hand, the ID expert centered cycle (left path in Fig. 1) concerns the post-processing of results derived from the data mining 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 data mining tasks, during a new knowledge extraction cycle.

3.2 Interpreting and Evaluating Extracted Knowledge

We suggest three ways to evaluate the association rules, inspired from a classification presented by Geng and Hamilton [10]: *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 is a traditional knowledge evaluation 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, and of the support and confidence of the obtained rules.

The minsup and minconf thresholds are pre-chosen for applying an extraction algorithm (here, the well-known Apriori algorithm [5]). Indeed, they provide 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 low minsup would lead to a combinatorial explosion of the number of candidate itemsets; on the contrary, a high minsup would prevent the appearance of association rules containing rare attributes [11], which are often interesting. If minsup=0, each considered transaction is 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. On the opposite, the rules with low confidence may be inconsistent, but may also express unusual but interesting situations.

In practice, an efficient processing of the attributes characterizing the transactions requires to test different thresholds, since rare rules are often more interesting than frequent ones.

In that context, many studies on association rules evaluation are limited to determine the interest of a rule from a statistical point of view, resulting in a lot of inconsistent rules, or just uninteresting ones from the point of view of the expert user.

We describe here an attempt to complete the classic rules evaluation in order to improve the quality of the results, in terms of quantity and quality. As a complement to "objective measures" (support-confidence), we suggest a semantic and subjective evaluation to create new and more relevant knowledge for the industrial domain expert user.

Semantic evaluation facilitates the evaluation of the interest of a rule according to the domain knowledge. 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)* analyse "elementary" rules (involving only two attributes), *ii)* express each analysed attribute by a question, *iii)* express the problem addressed by each rule by combining the questions, *iv)* interpret the support and confidence of rules, *v)* analyse the potential use of each rule for improving the industrial processes, *vi)* check 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 in the definition of the attributes), *vii)* analyse 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)* represent 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)* formalise 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 is related to looking for specific types of rules according to the user expectations (ID expert). Structuring the rules indeed facilitates a visual exploration and assists the expert in this validation step.

In our KDD process, the target knowledge is not predetermined during the application of the extraction algorithm, unlike others techniques constraining the number of items and/or determining what items are in the hypothesis or conclusion part. However, an ID expert user in a given situation has usually an idea on the type of rule that he/she expects, in relation with the decisions he/she has to make.

Mining algorithms like Apriori [5] allow to identify 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 of high interest, providing new knowledge to the user.

In the literature, different techniques are suggested to perform the subjective evaluation of extracted rules. A study of several techniques based on knowledge/user expectations has been detailed by Marinica [7]. Several formalisms may be used to represent knowledge for filtering rules: templates, beliefs, meta-rules, queries,

taxonomies and ontologies for instance. It is commonly accepted that a visual representation of association rules facilitates the interaction with the user and may for instance help to model a query (user expectation), allowing to filter the identified rules. A query Q relates to a rule skeleton, describing the a priori structure of the rules of interest for the user. A query/answering mechanism will look for "response" rules to sort a final set of potentially interesting rules.

More formally, let X be a set of extracted rules and Q a user query. Regarding the structure of the extracted rules, Liu et al. [12] suggests to distinguish four sets of potentially interesting rules:

- Conforming rules: an extracted rule $X_i \in X$ conforms with 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 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 hypothesis 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.

4 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 extracted from the SAP ERP Production Maintenance module is our starting point, containing several attributes (date, 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 the Apriori algorithm, and the post-processing phase considering the industrial domain expert in the interpretation and validation of results. Extracted knowledge was discussed with the domain expert by presenting him the first partial results of the KDD process. This steps allowed to improve the extraction process (for example, by not taking into account attributes of questionable interest).

For filtering the extracted rules, we have firstly empirically chosen minsup=20% and minconf=90%, 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 details the meaning of some rules, taking into 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: 21% of the interventions are preventive and performed every 6 months. Every intervention that has a semi-annual frequency concerns a preventive intervention (conf=1.0). However, preventive interventions may have other frequencies (since the reverse rule is absent).
- *Rule 2.* Production=0001 → Type of equipment=XXX sup=0.23 conf=0.97
  Question answered: link between "what site" and "what type of equipment".
  Interpretation: 23% of the interventions concern the type of equipment XXX on the production site 0001. 97% of the maintenance activities on this production site correspond to this type of equipment.
- *Rule 5.* Model=Booths → Production=0002 sup=0.35 conf=1.0
  Question answered: link between "what model" and "on which site".
  Interpretation: 35% of the 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).

The other rules have also been analysed and may be considered as variants of those six basic ones by means of the three logical operations (extension, permutation and junction). Therefore, we provided to the ID expert a model generalizing the extracted rules set (Fig. 2), specifying the logic relation between the elementary rules (upper part in Fig. 2) and more complex rules. For instance, R9 is an extension of R4 (new items have been added to the hypothesis part of R9), R9 is a permutation of R10 (some items of the rules R9 and R10 have been permuted), R10 is a junction between the rules R5 and R6 (combining the hypothesis parts of R5 and R6) and R10 is also a permutation of R8.

Finally, the domain expert may make 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 the proposed approach, the role played by the domain expert and the quality of the input data are decisive; both affect the quality of the extracted knowledge.

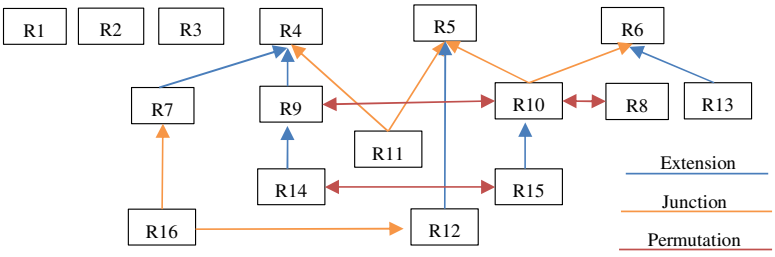

Fig. 2. Identification of the relationships between the elementary and complex rules

5 Conclusion

The interactive approach proposed for post-processing extracted association rules takes into account some efforts already reported in the literature; however, its main novelty is on the semantic interpretation and subjective evaluation of the extracted knowledge, according to several factors: support, confidence, presence and absence of expected rules, reverse rules, relationship between the extracted rules set and the frequent itemsets, and finally the interaction between the ID expert and the DM expert. The main focus is here on including considerations (positive and/or negative feedback) of the ID expert in order to improve the new knowledge extraction process in consistence with the application context. Indeed, it is essential to understand what the user is looking for in order to be able to define the problem and apply relevant data mining techniques. Other applications are in progress in the pharmaceutical and aeronautical domains, using more complex databases with more cases and attributes for improving and optimizing the interpretation and evaluation of extracted knowledge during the post-processing phase in the KDD process.

References

[1] Giudici, P.: Applied data mining: Statistical methods for business and industry. Wiley (2003)
[2] Fayyad, U.M., Piatetsky-Shapiro, G., Smyth, P., Uthurusamy, R.: Advances in knowledge discovery and data mining. MIT Press (1996)
[3] Harding, J.A., Shahbaz, M., Shahbaz, S., Kusiak, A.: Data mining in manufacturing: A review. Journal of Manufacturing Science and Engineering - Transactions of the ASME 128, 969–976 (2006)
[4] Köksal, G., Batmaz, I., Testik, M.C.: A review of data mining applications for quality improvement in manufacturing industry. Expert Systems with Applications 38(10), 13448–13467 (2011)
[5] Agrawal, R., Srikant, R.: Fast algorithms for mining association rules in large databases. In: Proceedings of the 20th International Conference on Very Large Data Bases, VLDB 1994, 1215th edn., pp. 487–499. Morgan Kaufmann Publishers Inc (1994)
[6] Choudhary, A.K., Harding, J.A., Tiwari, M.K.: Data mining in manufacturing: a review based on the kind of knowledge. Journal of Intelligent Manufacturing 20(5), 501–521 (2009)
[7] Marinica, C.: Association Rule Interactive Post-processing using Rule Schemas and Ontolo-gies-ARIPSO. PhD thesis, Ecole polytechnique de l'Université de Nantes (2010)
[8] Ben Ayed, M., Ltifi, H., Kolski, C., Alimi, A.M.: 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)
[9] Wang, H., Wang, S.: A knowledge management approach to data mining process for business intelligence. Industrial Management & Data Systems 108(5), 622–634 (2008)
[10] Geng, L., Hamilton, H.J.: Interestingness measures for data mining: A survey. ACM Computing Surveys 38(3), Article 9 (2006)
[11] Baesens, B., Viaene, S., Vanthienen, J.: Post-processing of association rules. DTEW Research Report 0020, pp. 1–18 (2000)
[12] Liu, B., Hsu, W., Wang, K., Chen, S.: Visually Aided Exploration of Interesting Association Rules. In: Zhong, N., Zhou, L. (eds.) PAKDD 1999. LNCS (LNAI), vol. 1574, pp. 380–389. Springer, Heidelberg (1999)