A User-driven and Quality-oriented Visualization for Mining Association Rules

Julien Blanchard, Fabrice Guillet, Henri Briand
IRIN – Polytech'Nantes – University of Nantes
La Chantrerie – BP50609
44306 Nantes cedex 3 France
{julien.blanchard, fabrice.guillet, henri.briand}@polytech.univ-nantes.fr

Abstract

On account of the enormous amounts of rules that can be produced by data mining algorithms, knowledge validation is one of the most problematic steps in an association rule discovery process. In order to find relevant knowledge for decision-making, the user needs to really rummage through the rules. Visualization can be very beneficial to support him/her in this task by improving the intelligibility of the large rule sets and enabling the user to navigate inside them. In this article, we propose to answer the association rule validation problem by designing a human-centered visualization method for the rule rummaging task. This new approach based on a specific rummaging model relies on rule interestingness measures and on interactive rule subset focusing and mining. We have implemented our representation by developing a first experimental prototype called ARVis.

1. Introduction

Among the knowledge models used in Knowledge Discovery in Databases (KDD), association rules [1] have become a major concept and received significant research attention. These rules are implicative tendencies of the form $X \rightarrow Y$ where X and Y are conjunctions of database items (boolean variables). One of most problematic steps in an association rule discovery process is the post-processing of the rules, i.e. the interpretation, evaluation and validation of the rules after their extraction. Indeed the data mining algorithms can produce enormous amounts of rules. In practice, it is very tedious for the user (a decision-maker specialized in the data studied) to find interesting knowledge for decision-making in a corpus that can hold hundreds of thousands of rules or even millions of rules with large business databases. This problem is due to the unsupervised nature of association rule discovery: the user does not make his/her goals explicit and does not specify any endogenous variable.

Three kinds of approaches aim at helping the user appropriate the bulks of association rules: reducing the number of rules with interestingness measures [13] or summary techniques [11], exploring the rules with interactive tools like rule browsers [9] or query languages [8], and visualizing the rule sets with visual representations like matrices or graphs [6, 14]. In this article, to apprehend the problem of rule validation, we have opted for defining the user's task as a prerequisite. Indeed in order to efficiently assist the user in his/her search for the interesting knowledge, the KDD process should be considered not from the point of view of the discovery algorithms but from that of the user's, as a user-centered and task-oriented decision support system [4]. From the definition of the user's task and the cognitive constraints which ensue, we propose an appropriate model of rule rummaging which follows from our previous works on the exploration of rule sets using graphs [10]. Then we present an interactive visualization method for the human-centered process of association rule rummaging. This method combines the three approaches described before with a tight association of interestingness measures, a strong interactivity with the user, and a visual representation. We have implemented our new visualization method in a first rule mining prototype called ARVis. Including an online algorithm of rule extraction, ARVis allows the user to mine the rules interactively *via* the visual representation all along the rummaging process.

In the next part we define the user's task and our model for the human-centered rummaging of association rules. Then we present our interactive visualization method and the choices we made for ARVis regarding the rule set structure, the visual metaphor, and the interactions.

2. Human-centered rummaging of association rules

2.1. User's task

During the knowledge validation step in the post-processing of the rules, the user is faced with the rules extracted by data mining algorithms and described by interestingness measures. The user's task is then to find interesting rules for decision-making in these long lists. Inspired by research works on the user's behavior in a knowledge discovery process [2] on the one hand, and

also by cognitive principles of information processing in the context of decision models [12] on the other hand, we consider that the user applies a focusing strategy to apprehend the bulk of rules: faced with a large amount of information, the user focuses his/her attention on a limited and therefore more intelligible subset of potentially useful information. To facilitate the user's post-processing task, we have developed a model of human-centered rummaging which supports his/her focusing strategy by allowing to isolate a rule subset, to explore it, and to change it in an iterative way until he/she is able to reach a decision.

2.2. Rule rummaging model

Our rule rummaging model consists in letting the user navigate as he/she wishes through the voluminous rule set by focusing on the successive limited subsets to explore. The user drives a series of local explorations by trial and error through the whole rule set from which only the selected portion is gradually visited. Implementing such a rule rummaging process implies structuring the rule set to allow the user to focus on rule subsets and navigate from one to another. More precisely, we need to group the rules together into subsets and combine these subsets among themselves by neighborhood relations (figure 1). At each navigation step, to access a new rule subset after exploring the current one, the user has the choice among all the neighboring subsets, i.e. those reachable by the neighborhood relation. This relation can be implemented by a retrieval procedure if the rule set is already extracted (post-analysis of rules), or by a local algorithm for constraint-based rule extraction (inductive database approach [7]) so that the user drives the data mining interactively from the rummaging process.

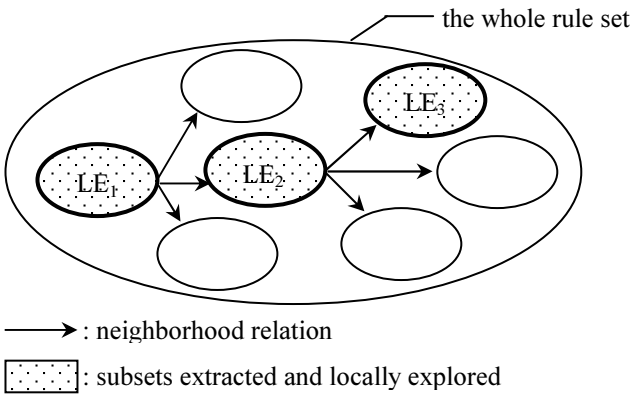

Figure 1. The neighborhood relation allows to mine the rule subsets according to the user's navigation

3. Visualization for rule rummaging in the ARVis tool

In this part, we describe the choices we made to implement the human-centered rummaging process in ARVis. The rules are here described by three interestingness measures: support, confidence [1], and implication intensity (respectively noted *sp*, *cf* and *ii*). Support evaluates the generality of the rules and confidence its validity (success rate), while implication intensity evaluates the rule statistical surprisingness by quantifying the unlikelihood of the number of counter-examples compared to a probabilistic model [3, 5]. Each measure is associated to a minimal threshold set by the user and exploited to filter the rules: s_{sp}, s_{cf}, s_{ii}.

3.1. Relation of specialization/generalization

Given the set I of items relative to the studied domain, the rules are of the form $X \rightarrow y$ where X is an itemset $X \subseteq I$ and y is an item $y \in I \setminus X$. We have chosen to structure the rule set by creating rule subsets *RULES(X)*, each corresponding to an itemset $X \subseteq I$. Each subset *RULES(X)* contains two kinds of rules, the specific ones *RULESspe(X)* and the general ones *RULESgen(X)*:

$$RULES(X) = RULESspe(X) \cup RULESgen(X).$$

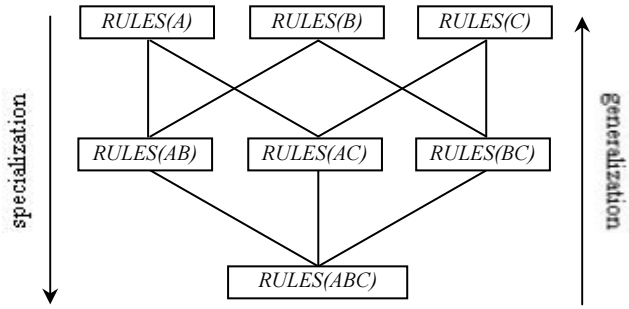

Figure 2. The relations of specialization and generalization among the rule subsets (with a set of items *I* = {*A*, *B*, *C*})

The specific and general rules are defined as:

$RULESspe(X) = \{ X \rightarrow y$ such as:
$y \in I \setminus X, sp(X \rightarrow y) \geq s_{sp}, cf(X \rightarrow y) \geq s_{cf}, ii(X \rightarrow y) \geq s_{ii} \}$

$RULESgen(X) = \{ X \setminus \{y\} \rightarrow y$ such as: $y \in X,$
$sp(X \setminus \{y\} \rightarrow y) \geq s_{sp}, cf(X \setminus \{y\} \rightarrow y) \geq s_{cf}, ii(X \setminus \{y\} \rightarrow y) \geq s_{ii} \}$

The specific rules have all the same left-hand side and only their right-hand sides differ, while the general rules are all built from the same items. We use as neighborhood relation a relation of specialization among the subsets and its symmetrical relation of generalization (see the graph figure 2). Specializing a subset *RULES(X)* amounts to

adding an item to the itemset X, whereas an item is removed by generalizing.

3.2. Quality-oriented visual metaphor

To generate a visual representation of the rules, we take advantage of the user's focusing strategy by only representing the current subset at each navigation step. Each rule subset is visualized using a 3D "information landscape" representation. With the use of 3D instead of 2D, the most important information can be displayed in the foreground, letting the less important information in the background. For each rule subset, the landscape is shared into two areas: one is dedicated to the specific rules, and the other one to the general rules.

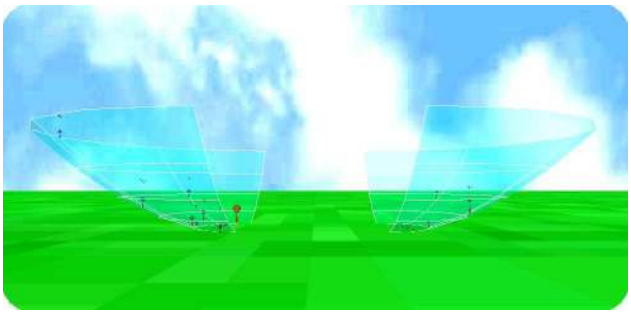

Figure 3. Two arenas in a landscape

We symbolize each rule by the following object: a sphere perched on top of a cone in the landscape. In the specific rule and the general rule areas, the objects are laid-out in the landscape on an arena (a "glass" half-bowl) to reduce occultation (figure 3). We have opted for the following visual metaphor to represent the rule subsets (figure 4):

- the object position represents the implication intensity,
- the sphere visible area represents the support,
- the cone height represents the confidence,
- the object color is used redundantly to represent a weighted average of the three measures.

This visual metaphor stresses the good rules whose visualization and access are made easier compared to the worse rules. Furthermore, some complementary text labels appear above each object to give the name of the corresponding rule and provide the numerical values for support, confidence and implication intensity.

3.3. Interactions

The user interacts in three different ways with the visual representation: by visiting the rule subsets, by filtering the rules on the interestingness measures, and by navigating among the subsets.

The user can wander freely over the 3D landscape to browse the rules and examine them more closely. To facilitate the user's exploration, there exist predefined viewpoints for the overall vision of each arena and for the close vision of each object.

During the subset local explorations, the user can filter the rules in the landscape by dynamic queries on the interestingness measures *via* sliders. These queries alter the thresholds s_{sp}, s_{cf}, s_{ii} on the measures to make only the rules with sufficient quality appear.

Finally, to drive his/her rummaging process, the user can navigate from one rule subset to another by clicking on the objects in the landscape. By clicking on a specific (respectively general) rule $X \rightarrow x$, he/she triggers the relation of specialization (resp. generalization), the current subset is thus replaced by the new more specific subset $RULES(X \cup \{x\})$ (resp. the new more general subset $RULES(X)$) and the representation is updated. Besides, the more specific or more general subsets the user can reach are represented in a shrunk version inside the spheres of the current landscape. This allows to anticipate whether a subset is worth exploring or not.

The relations of specialization/generalization are implemented using the hybrid mining algorithm presented in [10]. The first step of this algorithm is the "frequent itemset" mining procedure of the well-known *A Priori* algorithm [1]. The second step is an online local procedure polynomial in number of items which dynamically computes the rule subsets $RULES(X)$. Therefore, only the rules required by the user along his/her rummaging process are extracted.

4. Conclusion

In this paper, we have presented an interactive visualization method specially designed to support association rule mining and post-processing in a KDD process. This new approach is based on our human-centered model of rule rummaging appropriate to the user's task. It enables him/her to navigate through the voluminous rule set by carrying out a series of local explorations of limited subsets he/she focuses on. The user can thus comprehend the bulk of rules more easily to find relevant knowledge for decision-making. Coupled with an online algorithm of rule extraction, the visualization allows to interactively drive the data mining by producing only the rules required by the user. The subjectivity present in the rule validation step can thus be exploited in the KDD process to reduce the rule profusion. Moreover, our visualization takes advantage of the rules' names, used in the interactions to navigate among the subsets, but also of the interestingness measures, highlighted in the representation, which is original compared to the other rule set visualization methods.

We have developed ARVis, a first rule mining prototype implementing our visualization method. ARVis

is built on a client/server architecture. On the server side, a CGI program takes charge of the local extraction of the rules and of the visual representation construction in VRML. The rules and the 3D landscapes are therefore generated dynamically. On the client side, the user visualizes the landscapes with a web browser equipped with a relevant VRML plug-in. The tool can be used with shutterglasses to provide stereoscopic display in order to improve the perception of depth. Our future works will mainly concern:

- developing the tool for rule rummaging in Java3D,
- implementing additional neighborhood relations with appropriate local mining algorithms.

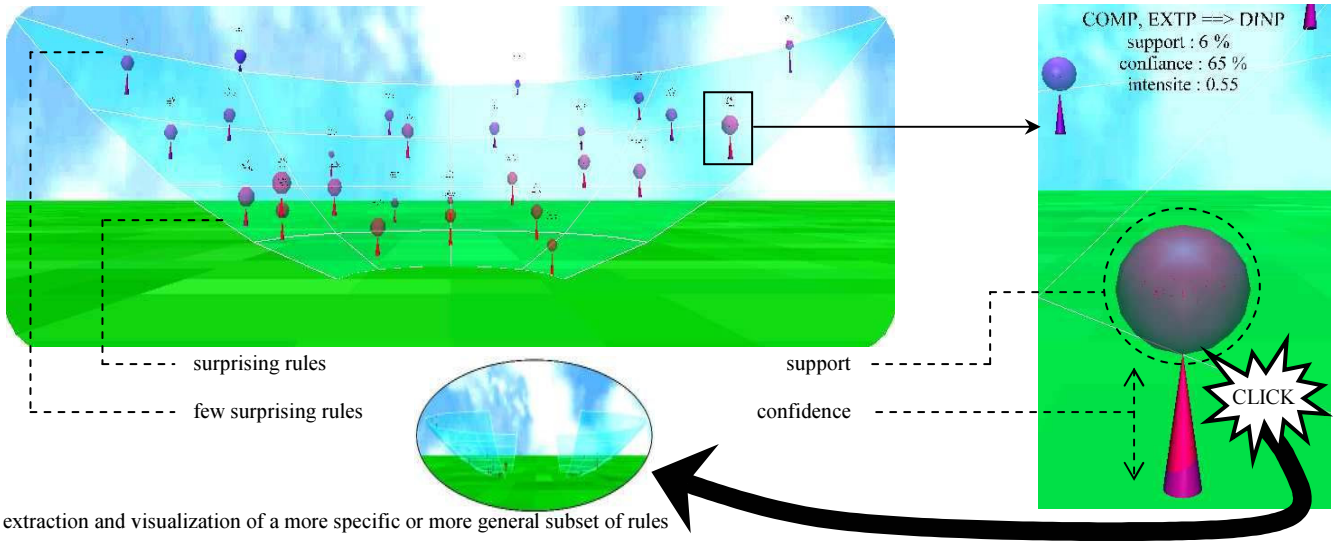

Figure 4. The visual metaphor and the interactions in ARVis

5. References

[1] Agrawal, R., Mannila, H., Srikant, R., Toivonen, H., and Verkamo, A.I., "Fast discovery of association rules", *Advances in Knowledge Discovery and Data Mining*, AAAI/MIT Press, U.M. Fayyad, G. Piatetsky-Shapiro, and P. Smyth (Eds.), 1996, 307-328.

[2] Bandhari, I., "Attribute focusing: machine-assisted knowledge discovery applied to software production process control", *Knowledge acquisition 6*, 1994, 271-294.

[3] Blanchard, J., Kuntz, P., Guillet, F., and Gras, R., "Implication intensity: from the basic statistical definition to the entropic version", *Statistical Data Mining and Knowledge Discovery*, CRC Press, H. Bozdogan (Ed.), 2003, 473-485.

[4] Brachman, J.R., and Anand, T., "The process of knowledge discovery in databases: a human-centered approach", *Advances in Knowledge Discovery and Data Mining*, AAAI/MIT Press, U.M. Fayyad, G. Piatetsky-Shapiro, and P. Smyth (Eds.), 1996, 37-58.

[5] Guillaume, S., Guillet, F., and Philippe, J., "Improving the discovery of association rules with intensity of implication", *Proc. of the 2nd European Conference of Principles of Data Mining and Knowledge Discovery*, Springer, L.N.A.I. 1510, 1998, 318-327.

[6] Han, J., Chiang, J., Chee, S., Chen, J., Cheng, S., Gong, W., Kamber, M., Koperski, K., Liu, G., Lu, Y., Stefanovic, N., Winstone, L., Xia, B., Zaiane, O.R., Zhang, S., and Zhu, H., "DBMiner: a system for data mining in relational databases and data warehouses", *Proc. of CASCON'97*, 1997, 249-260.

[7] Imielinski, T., and Mannila, H., "A database perspective on knowledge discovery", *Communications of the ACM 39(11)*, 1996, 58-64.

[8] Imielinski, T., and Virmani, A., "MSQL: a query language for database mining", *Journal of data mining and knowledge discovery 3(4)*, 1999, 373-408.

[9] Klemettinen, M., Mannila, H., and Toivonen, H., *Interactive exploration of discovered knowledge: a methodology for interaction, and usability studies*, Technical report C-1996-3, University of Helsinki, 1996.

[10] Kuntz, P., Guillet, F., Lehn, R., and Briand, H., "A user-driven process for mining association rules", *Proc. of the 4th European Conference of Principles of Data Mining and Knowledge Discovery*, Springer, L.N.A.I. 1910, 2000, 160-168.

[11] Liu, B., Hu, M., and Hsu, W., "Multi-level organization and summarization of the discovered rules", *Proc. of the 6th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2000, 208-217.

[12] Montgomery, H., "Decision rules and the search for dominance structure: toward a process model of decision-making", *Analyzing and Aiding Decision Processes*, P.C. Humphreys, O. Svenson, and A. Vari (Eds.), 1983, 471-483.

[13] Tan, P., Kumar, V., and Srivastava, J., "Selecting the right interestingness measure for association patterns", *Proc. of the 8th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2000, 32-41.

[14] Wong, P.C., Whitney, P., and Thomas, J., "Visualizing association rules for text mining", *Proc. of the IEEE Symposium on Information Visualization InfoVis'99*, 1999, 120-123.