Concept of Semantic Knowledge Base for Data Mining of Business Rules

Stanislav Vojíř

Department of Information and Knowledge Engineering,
Faculty of Informatics and Statistics
University of Economics, Prague, Czech Republic,
stanislav.vojir@vse.cz

Abstract. This paper presents a concept of a semantic knowledge base (in RDF form). This knowledge base will be usable for combination of data mining of association rules and definition of business rules sets. The data mining and domain experts will be able to use it for extension of possibilites of definition not only user-defined business rules, but also business rules generated from association rules gained from data mining tasks. This research is a part of EasyMiner project.

Keywords: data mining, association rules, business rules, knowledge base, background knowledge

1 Introduction

Let's consider a business analyst, who is working in a small bank. For example, let's call him David. David should prepare a classification model for assessing the creditworthiness of clients. He has at his disposal a dataset describing history of loans repaying in last 3 years. Applying data mining algorithms yields a classification model in form of rules. David wants to combine gained classification model with an older rule set with description of problematic clients. The older rule set was prepared by a domain expert one year ago and was successfully applied yet. Currently, David has a problem: The rule sets are based on different data dictionaries, with another names of attributes with another named groups of values. He would like to have a complex data dictionary for combination of rules from different sources.

Today, there is an increasing demand for decision support systems. The limitation of quality of each decision support system is quality and complexity of the used knowledge base. Suitable form of a knowledge base is a set of business rules. The advantage of knowledge base in a form of business rules is its modularity. The saved rules can be independently interpreted and evaluated. For combining of rules obtained from multiple resources, it is necessary to define all rules using one shared data dictionary (in business rules terminology, it is called "terms dictionary"). This paper presents a concept of knowledge base (in RDF form), which is suitable for support of data mining of classification business rules.

2 Related work

Business rules are simple, user-friendly rules presented in textual or graphical form. Concept of business rule is based on the premise: "Rules build on facts, and facts build on terms." Research of "business rules" generation from data mining results is a current, yet not very exposed, research topic.

To the best of our knowledge, the only publicly available software solution to business rule learning is *Rule learner* [6]. Rule learner is a part of *OpenRules Decision Management System*. The learned rules are presented in the form of decision tables in Microsoft Excel worksheets.

There are especially systems for interactive data mining of association rules and classification models. It is appropriate to mention systems *MIME Framework* [2] and *BigML* [1]. Another approach based on clustering algorithms in combination with ontologies is proposed in [5].

In EasyMiner project, there has been explored the possibilities of interactive preparation of sets of business rules for classification tasks. The business rules are generated from association rules obtained from data mining results. An advantage of using association rules is their great descriptiveness. It is possible to build really complex classification model with only small risk of overfitting. The test results of automatic generation of business rules from association rules and a demo of business rules editor were presented on the conference RuleML 2014. [3] [7] This paper follows the demo implementation of business rules editor and simple rules base in EasyMiner project. The designed knowledge base should allow combination of rules from different sources and their simple generation from data mining tasks. This approach should simplify the preparation of rule based classification models and also evaluation of rules novelty. For the structure of the designed knowledge base it is also used experiences of building background knowledge base for data mining tasks definitions in the form of XML files. [4]

3 Concept of semantic knowledge base

For definition of business rules it have to be used a complex terms dictionary. Suitable form of this dictionary is an ontology. The dictionary is shared by all rules saved in the knowledge base. A shared dictionary will also be a good medium for connection of results gained from more data mining tasks - especially in case of cooperation of more data mining and domain experts. For data mining, instructions for data preprocessing should also be shared using the knowledge base.

The designed knowledge base should be usable in connection with different datasets. For this purpose, it is necessary to use an appropriate level of abstraction. The basic entities of the knowledge base are *meta-attributes*. Meta-attribute is an abstraction of one property from the real world (age, sex, height, loan, rating etc.). Each meta-attribute can be defined in more *formats*. The meta-attributes with their formats are *basic terms* for definition of rules. The knowledge base can also be divided into two parts: 1) definitions of preprocessings (for data mining), 2) saved (business) rules. The structure of basic entities is presented on Figure 1.

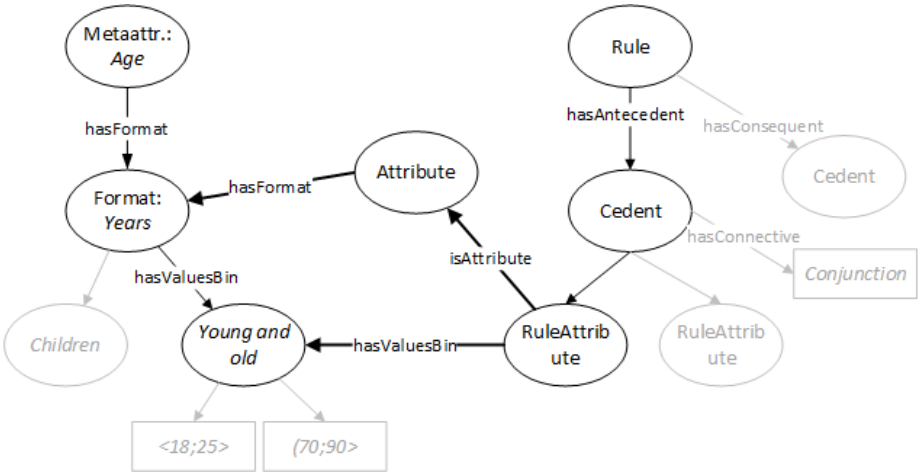

Fig. 1. Schema of meta-attributes and rules saved in knowledge base

How should it work: *An user wants to use EasyMiner for data mining analysis of a dataset and prepare a classification model. After uploading the dataset, EasyMiner system prompts the user to map columns from the dataset to meta-attributes (respectively their formats) saved in the knowledge base. For each column, the user selects a suitable mapping, or let the system to prepare a new meta-attribute/format. For data mining analysis, the user have to generate attributes from columns in the dataset (using preprocessing definitions). It is possible to simply use an existing preprocessing definition saved in knowledge base (for example "Age in years" can be divided into intervals with names "Children" and "Adult"), or define a new one. During the preparation of classification model, the user can select rules from data mining results, define own rules, or select rules from another tasks - without restrictions of attribute names and values bins. For example, the user can also define a rule about persons older than 70 years, but in the data mining attribute, there is only interval [18;+inf) called "Adult".*

The designed knowledge base should also be useful for evaluating the novelty of rules in data mining results. The user can view previously saved rules related (similar) to a selected rule and use them for rating of their similarity. The computational complexity of such request should not be too high. In the first step, the system will be able to filter rules based on same meta-attributes, in the second step, the overlaps of values bins should be evaluated. The knowledge base is saved in the form of RDF graph. The individual entities (e.g. meta-attributes, attributes, formats) are instances of ontological classes. They are identified by URIs. To create an applicable classification model, the selected rules are exported from the knowledge base to DRL form, which is used in *JBoss Drools*.[1]

[1] JBoss Drools - business rules management and execution system, `http://drools.org`

4 Conclusion and future work

Currently, the author has implemented a first version of the knowledge base presented in this paper. The knowledge base has been implemented in form of web application in PHP, which is accessible using REST API. It is currently used for integration of EasyMiner export of selected association rules with business rules editor and with component for testing of classification models. In the near future the knowledge base should replace currently used preprocessing module in the EasyMiner system. For more details, please visit `http://easyminer.eu`

The advantage of presented concept of knowledge base is in the direct connection of data mining tasks with knowledge base containing preprocessing definitions and with saved business rules. The assumption is that the connection of data mining datasets to knowledge base should greatly facilitate the data mining analysis and the preparation of classification models. The system should also be able to validate the novelty of founded data mining results.

The further work should be focused on testing of the described approach. There should be tested mainly the performance demands, especially in case of really big counts of triples saved in the RDF store. Possibly, it could be suitable to check more variants of structure of the knowledge base.[2] Following steps should be implementation of the validation for the novelty of rules in data mining results and the investigation of an effective way of values mappings between different formats.

References

1. BigML, Inc. Bigml is machine learning for everyone. `http://www.bigml.com`. Accessed: 2014-05-30.
2. Bart Goethals, Sandy Moens, and Jilles Vreeken. MIME: A framework for interactive visual pattern mining. In *Proceedings of the 17th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '11, pages 757–760, New York, NY, USA, 2011. ACM.
3. Tomáš Kliegr, Jaroslav Kuchař, Davide Sottara, and Stanislav Vojíř. Learning business rules with association rule classifiers. In *Rules on the Web. From Theory to Applications*, pages 236–250. Springer, 2014.
4. Tomáš Kliegr, Stanislav Vojíř, and Jan Rauch. Background knowledge and pmml: first considerations. In *Proceedings of the 2011 workshop on Predictive markup language modeling*, pages 54–62. ACM, 2011.
5. Claudia Marinica and Fabrice Guillet. Knowledge-based interactive postmining of association rules using ontologies. *IEEE Transactions on Knowledge and Data Engineering*, 22(6):784–797, 2010.
6. OpenRules, Inc. Rule learner. `http://www.openrules.com/rulelearner.htm`. Accessed: 2014-04-15.
7. Stanislav Vojıř, Přemysl Václav Duben, and Tomáš Kliegr. Business rule learning with interactive selection of association rules. In *RuleML Challenge 2014*. 2014.

[2] In the currently presented version, each rules consists of many entities, but it could be saved also in another structured form in one entity (e.g. XML or JSON).