J. Yaghob (Ed.): ITAT 2015 pp. 59–64
Charles University in Prague, Prague, 2015

Extension of Business Rule Sets Using Data Mining of GUHA Association Rules

Stanislav Vojíř
Department of Information and Knowledge Engineering
University of Economics, Prague
W. Churchill Sq. 4, Prague 3, 130 67, Czech Republic

Abstract. *The following paper is intended to introduce three suitable ways of using data mining of GUHA association rules in conjunction with existing set of business rules. The integration can be realized using full integration, as black box classification model and also using dynamic integration with data mining system. These ways are illustrated by demo use case based on data from a health insurance company.*

1 Introduction

Business rules are not only an effective way for modeling of business structure and descriptions of operations, definitions and constrains in an organization, but also an efficient way for separation of business logic from the application code of information systems. The separation of business logic, mainly "decision-making points" from the implementation of applications is very important, especially in today´s rapidly changing world. For this reason, it can be observed an increasing number of applications of rule engines and business rules system.

In this paper, the presented approach of extension of a business rules base is illustrated using examples from a health insurance company. From this domain, examples of business rule could be: *"If the doctor has specialization 001, then the diagnosis AAA is OK."* or *"The child emergency cannot treat the adult patients."* Such rules are usually saved and managed by a business rule management system. The rule set in conjunction with the related terms dictionary can be called "knowledge base".

However, the applicability of the rule-based systems greatly depends on the complexity and completeness of their knowledge base. In addition to the manual input of business rules by domain experts, there have been discovered also some methods of obtaining business rules from the business data – for example from unstructured texts or from operational data store of the company. A suitable method for "learning" of business rules from the working or historical business data is application of data mining methods and reusage of the gained data mining models.

1.1 Related Work

From the relevant works and papers, the "semi-automatic learning of business rules" has been a subject of research activities for relatively long period. But there are still not too many real applications. The most relevant existing application of "data mining of business rules" is the component *RuleLearner*, which is a part of the business rules system *OpenRules.*[1] This system works with knowledge base in the form of decision tables in Excel worksheets. According to the information from the company OpenRules, Inc., the component RuleLearner is still non-public. It is based on data mining using open source system Weka, but the conversion from data mining results to the form of classification tables for the OpenRules system can be realized only by experts from the authors´ company.

1.2 Business Rules

In this paper, the author describes three suitable ways of direct integration of data mining results into an existing business rule set. *Business rules* is not the name of one specification or system. The term "business rules" covers the relatively great area of rule-based systems and applications. It is mainly the name of modeling approach. In this approach, the modeling of the business behavior and decisions leads from the definition of basic entities and terms to the definition of standalone business rules. These rules are collected info rule sets in one complex knowledge base of the company.

The business rules approach has been applied in many specifications of languages for definition of business rules. The specifications can be divided by their main focus in two groups – specifications suitable for inference engines and specifications suitable for sharing of knowledge in human-friendly form. The work presented in this paper is more suitable for implementation in automatic inference (business rules) engines – JBoss Drools, Jess, Jena etc. The execution component takes the set of business rules and the base of facts, evaluates the conditions of business rules and activates the proprietary rules.

1.3 GUHA Association Rules

One of the possible and suitable methods for extension of knowledge base in the form of business rules is the application of data mining methods on the historical data of the company. It seems that the suitable data mining models are association and decision rules. The association rules can be discovered not only using the mostly known algorithm *APRIORI*, but also using the procedure *ASSOC* of the *GUHA method*.[1]

The GUHA method is original Czech data mining method for data mining of association rules with "rich semantic". The basic form of GUHA association rules is

$$\varphi \approx \psi$$

where φ *(antecedent),* ψ *(consequent*[2]*)* and possibly are logical combinations of attributes (with concrete values) and ≈ is the quantifier – function defined on the four feet table. Examples of the 4ft-quantifiers are *founded*

[1] In this paper, the rules founded using application of GUHA procedure ASSOC are called „GUHA association rules".

[2] In the GUHA method, *consequent* is called *succedent*

implication (combination of interest measures *confidence* and *support*) and *above average dependence* (this quantifier is convertible to the combination of interest measures *lift* and *support*). [5]

The GUHA association rules for the approaches presented in this paper are discovered using the data mining system LISp-Miner.[3] This software supports data mining of GUHA association rules also with the "dynamic binning of values in attributes". This feature extends the pattern of requested association rules (task definition). The attributes can contain the set of values – for example the rule attribute *age([0;1),[1;5))* is interpreted as age in interval from 0 to 5 years (without the request for redefinition of the data preprocessing). The dynamic binning can be defined as subsets of the given length, left or right cuts, intervals etc.

An example of the founded GUHA association rule:

age([20;40]) & city(Prague) & clinic(A, B) →
procedure(C) | confidence 0.6, support 0.01

The interpretation of this rule: If the age is in the interval from 20 to 30 years, city is Prague and the clinic is A or B, then the applied procedure is C. The confidence of this rule is 60% and support is 1%.

1.4 Structure of this Paper

This work is focused on the use of association rules obtained by application of *GUHA method* (below in text called "GUHA association rules"), but the principles are generalizable also for the usage of simpler association rules obtained using the algorithm APRIORI (for example in the system R). This paper follows the previous work of preparation classification business rule sets using GUHA association rules [2] and is also related to currently solved TAČR project TA04011691 "Automated extraction of business rules with feedback" [3].

The paper is organized as follows. Section 2 gives a walk through three suitable models of integration data mining model into business rule set. Section 3 contains example use cases motivated by real data. The conclusion summarizes the paper and outline for future work.

2 Integration of Data Mining Models into Existing Business Rule Set

Within this section, there are described three model ways of integration GUHA association rules into an existing business rule set. The suitability of their use differs according to the requested level of the integration and also to the analytical questing solved with the data mining task. All these ways are fully implementable (and have been practically verified) using business rule engine JBoss Drools [4] and data mining system LISp-Miner [5].

2.1 Direct Ttransformation of GUHA Association Rules into Business Rules

First variant of the involvement of founded association rules into an existing business rule set is the *direct transformation of them*. Within this transformation, every founded GUHA association rule is transformed into a separate business rule. From the GUHA association rule, antecedent and condition parts are transformed into condition of the business rule, consequent[4] of the association rule is "implemented" in the body of the business rule. The body of the business rule executes the requested action – returns the result of the classification task in suitable form (set of attributes with values, adds new data in the base of facts etc.) For this transformation, some constrains of the solved data mining tasks has to been considered.

Antecedent, condition and even consequent of a GUHA association rule can consists from multiple "partial cedents" (brackets in logical representation), containing conjunctions, disjunctions and negations. In case of mining using LISp-Miner system, every attribute in the rule can also contain multiple values, connected during the mining process using the "dynamic binning" feature. For the possibility of transformation from association rules to business rules, it is not necessary to apply any limits or constrains to antecedent and condition part of association rules. However, it is necessary to solve the *problem of the data dictionary*. The data dictionary has to be mapped to shared terms dictionary used in organization. If the data mining process has been initialized using data from operational data store of the organization, it is possible to use the default names of data attributes (columns) in the operational data store as the terms dictionary for definition of business rules.[5]

From the perspective of transformation to the form of business rules for the system JBoss Drools, condition of the rule can consist from logical expressions similar to native java code. The transformation consists from these steps:

1. Perform *reverse preprocessing* of used data. In data mining process it is common to prepare attributes from the data columns from the original data matrix. These attributes have different names and preprocessed values (during the preprocessing phase of data mining process, the original data values are grouped into named sets or intervals of original data values). The transformation itemizes the attributes included in association rules to the original names and values.
2. Remove unnecessary cedents from antecedent and condition part of GUHA association rule – because of the data mining task configuration and LISp-Miner export, the GUHA association rules saved in PMML[6] form often contain unnecessary partial cedents (multiple brackets without any added logical expression).
3. Transform antecedent and condition of every GUHA association rule into condition of a business rule. Dependently on the handling method of null values in the data set for data mining task, negation in association rule can be interpreted as *inequality* or

[3] http://lisp-miner.vse.cz

[4] In GUHA method is „consequent" called „succedent".

[5] Alternativelly in the organization maybe exists a mapping for data attributes from operational data store to an ontology or other "terms dictionary".

[6] Predictive Model Markup Language – XML-based format (technical standard) for saving of data mining models; developer by Data Mining Group

negation of the checking condition. For preparation of a classification business rule set, it is more suitable to use the interpretation as inequality (by testing results). Negation in association rule expression should be interpreted as *inequality*. In case of mining of GUHA association rules with condition, the condition can be appended to antecedent part (using conjunction), or could be interpreted as group condition for conditioned subset of business rules.

4. Prepare business rules´ bodies from the consequents association rules cedents. Semiautomatic acquisition of business rules from data mining results is suitable for solving of "classification" tasks. These tasks cannot return value of one "result" attribute. The limitations of consequent of the association rules for following automatic processing of results are as follows: Each consequent should contain one or more attributes with values, which were not preprocessed in data mining process. In case of more attributes in consequent part of association rule, these attributes should be connected within conjunction.
5. Use requested conflict resolution strategy.

Business rules in DRL form (format suitable for JBoss Drools) are based on Java classes, which represents the terminological dictionary. For support of solving classification problems using association rules, in most cases it is necessary to select the best result consequent (the resulting recommendation) in case of more business rules with matching antecedent/condition. Good conflict resolution strategy is to prefer classification rules with better values of confidence, support and shorter condition.[6] In DRL, the suitable strategy is implemented in one conflict resolution function written in DRL.

The result of recommendation/classification task can be processed with other part of information system of the organization, or can be processed with other business rules. Based on testing use cases, it can be said, that the following processing of the results using other business rules contributes to the clarity of the full knowledge base of the organization. From the perspective of knowledge management in organization in context of business rules, it is appropriate to build one shared knowledge base in form of business rules based on one shared terms dictionary. [7]

In implementation using JBoss Drools, it is suitable to (temporarily) insert results of classification subtask into the base of facts and continue in the business rules execution.

Great advantage of the transformation of each one association rules into a separate business rules is the possibility of their subsequent management and administration using tools from the business rules management system. It is easy to edit these rules, their priority and behavior.

In case of automatic transfer of the complete results of data mining of GUHA association rules into business rules, there can be also found some disadvantages. First big disadvantage of full integration is a large increase of the number of business rules. For solving of classification tasks using association rules without pruning algorithms, it is suitable to use data mining tasks with a really low requested minimal threshold value of support. Such tasks, however, return a lot of founded rules (possibly thousands of rules). In case of their integration into the main knowledge base, it is appropriate to identify these rules with specific "tag".

In terms of practical evaluation, the options of this model of integration were verified in [2] and [8]. It is suitable to generate business rule set in DRL form from GUHA association rules. The classifier obtained by this method can achieve even better results than reference classifiers. [2] According to realized tests, dependently on the solved data set, the greater "expression language" of GUHA association rules can contribute to better results (but at the cost of more rules).

2.2 Black Box Classification Component

The second suitable variant for the inclusion of data mining results into an existing knowledge base in form of business rule set is the integration as "black box". In this way, the connected component is suitable for solving of classification tasks. The integration schema should be as follows:

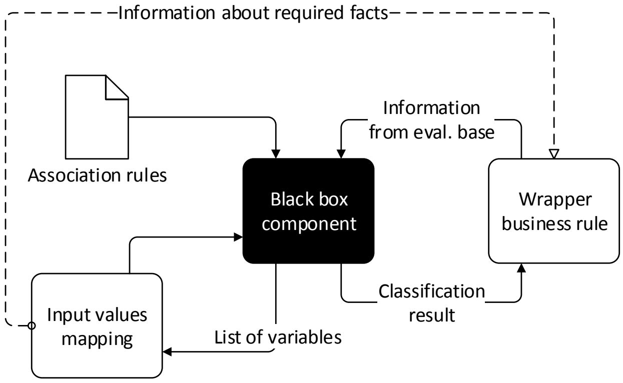

Fig. 1. Schema of black box component integration.

The user connects the black box component as one "part" of the knowledge base. It can be connected to body of a business rule, or as a partial condition. In the definition phase of the issued business rule, the user has to follow steps of a simple connection process:

1. Select results of a data mining task and export them into a standardized form (usually PMML).
2. Define *wrapper rule* – one rule, which initializes the evaluation of a classification black box component.
3. Import data mining results into the classification black box component. The component checks the structure of the uploaded model and detects all connecting points. The connecting points could be defined as input, output or shared. Input connecting points should include a definition of mapping between facts in the evaluation base and attributes used in conditions of the classification model.
4. Define mapping for the connecting points: In case of classification model based on GUHA association rules, the user defines 1:1 mapping between attributes used in antecedents and conditions of association rules and fields from the terms dictionary, the output connecting point is usually a variable for the result of

classification. The result variable could be immediately captured and processed in the wrapper business rule, or added into the evaluation base of facts used in the inference algorithm. For all the mappings, the black box component detects required data types for individual attributes and checks the mapping at least on the level of data type, at best on level of the definition range.

The involvement of a data mining model as the black box component brings many benefits. This way of integration has the lowest requirements for interaction with other rules in the knowledge base and it is applicable not only for data mining models consisting of rules but also for other suitable types of data mining models. For example, there can be considered decision trees or neural networks, too.

From the perspective of management or domain experts, this integration does not have too big impact on other business rules saved in the knowledge base. It is really easy interpretable: "In the condition of this rule matches the characteristic of client, the body of the rule returns the statistically most probable next offer for the client." The "most probable next offer" is determined with the black box component, so the management expert does not have to know the hidden algorithms used for this recommendation.

This integration has also disadvantages. The most of them is the problematic of "recycling" of specified data mining models for usage in more business rules. The data mining model is usually connected at only one point (in the black box component integrated in wrapper rule"). In case of usage models based on rules is a disadvantage also the exclusion of the evaluation of contained rules out of the main RETE network.[7]

In case of implementation of the black box component in the system JBoss Drools, it is possible to use external implementation in Java code, or implementation using separated, conditioned subset of business rules, which is evaluated only "on demand" (separated with special condition).

This model of integration has recently been implemented in TAČR project mentioned in Introduction of this paper.

2.3 Data Mining Initialized by Business Rules

Although the use of data mining models for solving of classification tasks integrated in business rule set is appropriately interpretable and user comprehendible, it is not suitable to limit the possible use cases for using only this way. The main reason for finding other, alternative approach is absence of the target attribute for classification in the operational data of the organization. Particularly in the case of usage data mining methods for finding of exceptions it is suitable to use dynamic data mining initialized by business rules. This process of definition the appropriate wrapper for initialization of data mining thought business rules engine in combination with LISp-Miner system could be defined as follows:

1. Define export from the operational data store of the organization. This export can be realized for example using SQL query and should be "repeatable" for later usages. The best way is definition of a view.
2. Define data mining task for selection of GUHA association rules in the data mining system LISp-Miner. Execute the task and check the results for the corresponding form. There are no limits for definition of the data mining task except of the "*final attribute*", which should be returned as result. This attribute should contain values from the original data matrix (without the use of values grouping in preprocessing phase or dynamic binning).
3. Export definition of the data mining task in PMML.
4. Define the wrapper business rule including the definition of data mining task, mapping of terms dictionary at least for "final attribute", database connection string and limits for counts of requested results.

   For some use cases, it is possible to map not only the final attribute, but also another attributes with fixed value for the definition of a condition.
5. Define period or condition for activation of the defined wrapper business rule. Within implementation using JBoss Drools, both these options are possible.

The wrapper business rule initializes the execution of data mining task. It is possible to run the LISp-Miner system not only from the graphical user interface, but also from the command line. After receiving the results from the data mining system, the wrapper rule compares the count of founded association rules. If the count is within the requested interval, the wrapper business rule extracts values of the final attribute in the founded rules and adds them as new facts in the evaluation base for processing using other business rules.

In case of inappropriate count of founded data mining results, the wrapper business rule can reinitialize the data mining task with modified thresholds of interest measures. To find association rules the user usually defines thresholds of two interest measures (usually confidence and support, for some cases also lift and support).[8] If the system founds too many rules, it is possible to increase the minimal requested thresholds of interest measures and execute the data mining task again.

This method of integration is suitable for interaction between business rules saved in the knowledge base and data mining systems for detection of exceptions in the operational data. Whether the exception can be negative or positive. For example detection of an increase in staff performance. The advantage of the application of data mining methods is the better performance than in case of evaluation the data matrix using set of specific business rules. However, also a disadvantage has to be considered. The separation of statistical evaluation of the operational data matrix from the knowledge base for execution using

[7] Most systems for execution of business rules are based on usage of RETE algorithm, which allows quickly inference evaluation.

[8] In the GUHA method, confidence and support are included in 4ft quaintifier „Founded implication“, lift is compatible with „Above average dependance“ quantifier (AAD).

external system can be founded either as advantage or disadvantage. It depends on the specialization of the domain expert. From the point of view of marketing or business specialists, it will be probably evaluated as an advantage – it is really simplification of the knowledge base.

2.4 Terms Dictionary for Definition of Business Rules

For a definition of business rules, it is required to use a *terms dictionary*. This terms dictionary should contain declaration of basic entities used in the organization. These terms composites to *facts*, and facts composites into *business rules*. For expanding of a business rule set using data mining results, the “good” terms dictionary can be the schema of the main operational database used in the organization.

The mapping techniques are not subject of this paper. For the integration of data mining results into existing business rule set, the best way is a definition of mapping in the mode 1:1 not only at level of data attributes, but also at level of their values.

In using of business rules, the mapping can be realized on basis of usage of specific *mapping rules*. In JBoss Drools, it is possible to define rules with conditional validity. So if the “mapping rule” detects in the evaluation base, it adds one or more other facts (instances of Java object) representing the mapped fact. The added fact is present and valid only while the mapping rule is active (it´s condition is evaluated as true).

3 Demo use case

For better illustration of the appropriateness of ways of integrating data mining results into a business rule set, it is suitable to explain them on a demo use case. In this paper, the author represents them on use cases defined on data from a health insurance company.

In every insurance company, it is necessary to collect the most possible data from the real life and reuse them for the risks analysis and for detection of fraud techniques. In the domain of health insurance, the medical facilities send lists of performed procedures and request a financial compensation for them. Every request composites from identification of the medical facility, the concrete medical worker, identification of the patient and details about the diagnosis and performed procedures. After composition in one “data row”, respectively one data matrix, there are tens of data attributes.

The health insurance company has contracts with individual medical facilities, but it does not mean, that every facility requests only really performed procedures. The reason may be a mistake, of course, but also attempt to fraudulently acquire some finances. The insurance company should have a list of rules (optionally a knowledge base in form of business rules) for detection of patently false requests. For example if a family doctor requests finance for a surgical operation. But is it necessary to detect not only obvious errors in requests. The insurance company want to detect also unusual growths of performed procedures, which could be potentially evaluated as untruthful.

3.1 Direct acquisition of business rules

Most business rules for evaluation the correctness of requests from medical facilities is inputted manually by domain experts. For founding of unobvious relations in data, it is suitable to use data mining methods. The user can select some founded association rules, convert them into business rules and use them for following manual editing of the knowledge base.

To use data mining techniques, it is necessary to have access to the archive with operational data received in the past. In terms of medical procedures it is also necessary to respect the specificities of different seasons and impact of weather. For example, there are differences in frequency and types of illnesses and injuries between the summer and the winter.

On the basis of this data mining analysis, it is also possible to detect potentially interesting areas for application of models for automatic learning of business rules.

3.2 Classification model learning

Suitable analytical question for processing of the incoming data is the detection of facilities, which require probably too much procedures or unusual combinations of them. A concrete example could be redundant performing of laboratory analysis of blood or automatically request for RTG for all patients of a surgery. These unnecessary procedures are no benefit not only for the insurance company, but also for the patients.

To solve this task, it is possible to use historical data about the checks previously made in medical facilities in combination with results from these tasks. Based on these data, it is possible to prepare a classification model for recommending suitable facilities for the future check.

The classification model can be included into the knowledge base as native business rules, or better in form of black box component. The advantage of separated black box component is the simpler replacement of the full classification model with a newer version.

3.3 Periodically solved data mining task

Another interesting task suitably solvable using data mining methods is detection of unusual increase or decrease of performed medical procedures in a concrete medical facility compared to other facilities of the same type. This task cannot be resolved in the “flow check” system, but it is possible to solve it using archive of the incoming data.

It is suitable use case for application of periodical solving of a predefined data mining task. The domain experts defines a data mining task for founding GUHA association rules for example in form:

diagnosis(A) & facility(*) → procedure(B) / clinicType(A)

where *clinicType(A)* is condition of founded rules, the task is defined using AAD quantifier (interest measures are *lift* and *support*) and the expert want to process as results the values of the attribute *facility*. The expert defines interval of minimal threshold of interest measures and maximal count of requested rules.

The data mining is then executed periodically once per month and the business rules system initializes the request for the check in the indicated medical facilities.

4 Conclusion and future work

In this paper, the author presented three suitable ways of integration data mining results (mainly GUHA association rules) into a knowledge base in form of business rules, which are suitable for automatically execution. These models are applicable not only in conjunction with JBoss Drools system, they are generally applicable with all “execution oriented” business rules systems. For example, there can be mentioned systems Jess, Jena or ERIAN.

Within the further work, it is necessary to propagate methods of automatic integration of data mining results into business rule sets. Another task is finalization of a model of knowledge base for combination data mining tasks with definitions of business rules. The demo implementation of the knowledge base, which concept was presented in [9], should be extended to a public methodology.

Acknowledgment

This paper was processed with contribution of long term institutional support of research activities and by IGA project 20/2013 by Faculty of Informatics and Statistics, University of Economics, Prague.

References

[1] OpenRules, Inc., “Rule Learner,” *Open Rules* [online] http://openrules.com/rulelearner.htm [cit. 2015-01-28]

[2] Kliegr, T., Kuchař, J., Sottara, D., Vojíř, S.: Learning business rules with association rule classifiers. Rules on the Web. From Theory to Applications, Springer, 2014, 236-250

[3] Vysoká škola ekonomická v Praze and KOMIX s.r.o., TA04011691 - Automatizovaná extrakce byznys pravidel se zpětnou vazbou (2014-2016, TA0/TA), 2013

[4] Red Hat, Inc, “Drools,” Drools - Business Rules Management System (Java™, Open Source) [online] http://www.drools.org/, [cit. 2015-04-21]

[5] Rauch, J., Šimůnek, M.: Dobývání znalostí z databází LISp-Miner a GUHA. Oeconomica Praha, 2015

[6] Thabtah, F. A.: A review of associative classification mining. Knowledge Engineering Review **22(1)** (2007),. 37-65

[7] Ross, R.G.: Principles of the Business Rule Approach. Addison-Wesley Professional, 2003

[8] Vojíř, S., Kliegr, T., Hazucha, A., Škrabal, R., Šimůnek, M.: Transforming association rules to business rules: EasyMiner meets Drools. RuleML Challenge 2013, CEUR-WS.org, vol. 1004, 2013

[9] Vojíř, S.: Concept of semantic knowledge base for data mining of business rules. Znalosti 2014 Exhibice, Edukace a nacházení Expertů - Exhibition, Education and Expert finding. Praha: KIZI FIS, 2014, 132-136