Formal Frame for Data Mining with Association Rules – a Tool for Workflow Planning

Jan Rauch and **Milan Šimůnek** [1]

1 INTRODUCTION

The goal of this extended abstract is to contribute to the forum for research on construction of data mining workflows. We briefly introduce a formal framework called *FOFRADAR* (FOrmal FRAmework for Data mining with Association Rules) and then we outline how it can be used to control a workflow of data mining with association rules. We consider this relevant to associative classifiers that use association rule mining in the training phase [3].

We deal with association rules $\varphi \approx \psi$ where φ and ψ are general Boolean attributes derived from columns of analyzed data matrices. Symbol $\approx$ is called 4ft-quantifier and it stands for a condition concerning a contingency table of φ and ψ [6]. Such rules are more general than rules introduced in [1]. We consider data mining process as described by the well known CRISP-DM methodology.

The FOFRADAR is introduced in [5]. Its goal is to formally describe a data mining process such that domain knowledge can be used both in formulation of reasonable analytical questions and in interpretation of resulting set of association rules. No similar approach to dealing with domain knowledge in data mining is known to the authors. An application of the FOFRADAR in data mining workflows is outlined here for the first time.

2 FOFRADAR

FOFRADAR is based on a **logical calculus $\mathcal{LC}$ of association rules**. Formulas of $\mathcal{LC}$ correspond to the association rules $\varphi \approx \phi$ [4]. Such rules are evaluated in data matrices rows of which correspond to observed objects $o_1, \ldots, o_n$ and columns correspond to observed attributes $A_1, \ldots, A_K$. We assume that A_i has a finite number $t_i \geq 2$ of possible values $1, \ldots, t_i$ (i.e. *categories*) and $A_i(o_j)$ is a value of A_i in row o_j for $i = 1, \ldots, K$ and $j = 1, \ldots, n$.

Boolean attributes φ, ϕ are derived from basic Boolean attributes i.e expressions $A_i(\alpha)$ where $\alpha \subset \{1, \ldots, t_i\}$. A basic Boolean attribute $A_i(\alpha)$ is true in a row o_j of a given data matrix $\mathcal{M}$ if $A_i(o_j) \in \alpha$, otherwise it is false. Thus, we do not deal only with Boolean attributes - conjunctions of attribute-value pairs $A_i(a)$ where $a \in \{1, \ldots, t_i\}$ but we use general Boolean attributes derived by connectives $\wedge, \vee, \neg$ from columns of a given data matrix.

The *4ft-table* $4ft(\varphi,\psi, \mathcal{M})$ of φ and ψ in a data matrix $\mathcal{M}$ is a quadruple $\langle a, b, c, d \rangle$ where a is the number of rows of $\mathcal{M}$ satisfying both φ and ψ, b is the number of rows satisfying φ and not satisfying ψ, c is the number of rows not satisfying φ and satisfying ψ and d is the number of rows satisfying neither φ nor ψ. A $\{0, 1\}$-valued associated function $F_{\approx}(a, b, c, d)$ is defined for each 4ft-quantifier $\approx$. The rule $\varphi \approx \psi$ is true in a data matrix $\mathcal{M}$ if $F_{\approx}(a, b, c, d) = 1$ where $\langle a, b, c, d \rangle = 4ft(\varphi,\psi, \mathcal{M})$, otherwise it is false in $\mathcal{M}$.

Expression $A_1(1, 2, 3) \vee A_2(4, 6) \Rightarrow_{p,B} A_3(8, 9) \wedge A_4(1)$ is an example of association rule, $\Rightarrow_{p,B}$ is a 4ft-quantifier of founded implication. It is $F_{\Rightarrow_{p,B}}(a, b, c, d) = 1$ if and only if $\frac{a}{a+b} \geq p \wedge a \geq B$ [2]. There are various additional 4ft-quantifiers defined in [2, 4].

A deduction rule $\frac{\varphi \approx \psi}{\varphi' \approx \psi'}$ is correct if the following is true for each data matrix $\mathcal{M}$: *if* $\varphi \approx \psi$ *is true in* $\mathcal{M}$ *then also* $\varphi' \approx \psi'$ *is true in* $\mathcal{M}$. There are reasonable criteria making possible to decide if $\frac{\varphi \approx \psi}{\varphi' \approx \psi'}$ is a correct deduction rule [4].

FOFRADAR consists of a logical calculus $\mathcal{LC}$ of association rules and of several mutually related languages and procedures used to formalize both items of domain knowledge and important steps in the data mining process. They are shortly introduced below, relations of some of them to the CRISP-DM are sketched in Fig. 1.

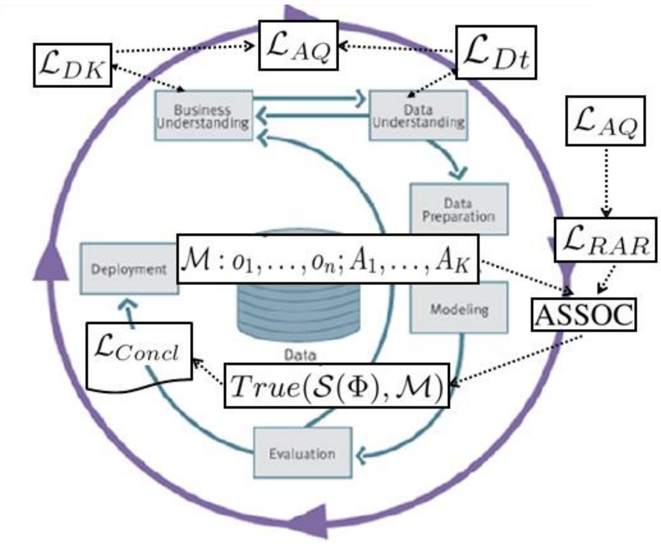

Figure 1. FOFRADAR framework and CRISP-DM methodology

Language $\mathcal{L}_{DK}$ ***of domain knowledge*** – formulas of $\mathcal{L}_{DK}$ correspond to items of domain knowledge. A formula $A_1 \uparrow\uparrow A_{11}$ meaning that if A_1 increases then A_{11} increases too is an example. We consider formulas of $\mathcal{L}_{DK}$ as results of business understanding.

Language $\mathcal{L}_{Dt}$ **of data knowledge** – its formulas can be considered as results of data understanding. An example is information that 90 per cent of observed patients are men.

Language $\mathcal{L}_{AQ}$ **of analytical questions** – the expression

[1] University of Economics, Prague, Czech Republic, email: rauch@vse.cz and simunek@vse.cz

$[\mathcal{M} : A_1, \dots, A_{10} \approx^? A_{11}, \dots, A_{20}; \not\to A_1 \uparrow\uparrow A_{11}]$ is an example of a formula of $\mathcal{L}_{AQ}$. It corresponds to a question $\mathcal{Q}_1$: *In data matrix $\mathcal{M}$, are there any relations between combinations of values of attributes $A_1, \dots, A_{10}$ and combinations of values of attributes $A_{11}, \dots, A_{20}$ which are not consequences of $A_1 \uparrow\uparrow A_{11}$?*

Language $\mathcal{L}_{RAR}$ **of sets of relevant association rules** – each formula Φ of $\mathcal{L}_{RAR}$ defines a set $\mathcal{S}(\Phi)$ of rules relevant to a given analytical question. The set $\mathcal{S}(\Phi)$ relevant to $\mathcal{Q}_1$ can consist of rules $\varphi \Rightarrow_{0.9,100} \psi$ where φ is a conjunction of some of basic Boolean attributes $A_1(\alpha_1), \dots, A_{10}(\alpha_{10})$, similarly for ψ and $A_{11}, \dots, A_{20}$. Here α_1 can be e.g. any interval of maximal 3 consecutive categories, similarly for additional basic Boolean attributes.

Procedure ASSOC – its input consists of a formula Φ of $\mathcal{L}_{RAR}$ and of an analyzed data matrix $\mathcal{M}$. Output of the ASSOC procedure is a set $True(\mathcal{S}(\Phi), \mathcal{M})$ of all rules $\varphi \approx \psi$ belonging to $\mathcal{S}(\Phi)$ which are true in $\mathcal{M}$. The procedure 4ft-Miner [6] is an implementation of ASSOC. It deals with a very sophisticated language $\mathcal{L}_{RAR}$.

Procedure Cons – this procedure maps a formula Ω of $\mathcal{L}_{DK}$ to a set $Cons(\Omega, \approx)$ of association rules $\varphi \approx \psi$ which can be considered as consequences of Ω. The set $Cons(A_1 \uparrow\uparrow A_{11}, \Rightarrow_{p,B})$ is a set of all rules $\varphi \Rightarrow_{p,B} \phi$ for which $\frac{\omega \Rightarrow_{p,B} \tau}{\varphi \Rightarrow_{p,B} \phi}$ is a correct deduction rule and $\omega \Rightarrow_{p,B} \tau$ is an atomic consequence of $Cons(A_1 \uparrow\uparrow A_{11})$. Rules $A_1(\mathit{low}) \Rightarrow_{p,B} A_{11}(\mathit{low}))$ and $A_1(\mathit{high}) \Rightarrow_{p,B} A_{11}(\mathit{high})$ are examples of atomic consequences of $A_1 \uparrow\uparrow A_{11}$, *low* and *high* are suitable subsets of categories of A_1 and A_{11}. Some additional rules can also be considered as belonging to $Cons(A_1 \uparrow\uparrow A_{11}, \Rightarrow_{p,B})$, see [5] for details.

Language $\mathcal{L}_{Concl}$ – formulas of this language correspond to conclusions which can be made on the basis of the set $True(\mathcal{S}(\Phi), \mathcal{M})$ produced by the ASSOC procedure. Two examples of such conclusions follow. (1): *All rules in $True(\mathcal{S}(\Phi), \mathcal{M})$ can be considered as consequences of known items of domain knowledge $A_1 \uparrow\uparrow A_{11}$ or $A_2 \uparrow\uparrow A_{19}$.* (2): *Lot of rules from $True(\mathcal{S}(\Phi), \mathcal{M})$ can be considered as consequences of yet unknown item of domain knowledge $A_9 \uparrow\uparrow A_{17}$.*

There are additional procedures belonging to FOFRADAR, they transform formulas of a particular language to formulas of another language of FOFRADAR [5].

3 FOFRADAR and Workflow of Data Mining

To keep things simple and the explanation concise we assume that the analyzed data matrix $\mathcal{M}$ is given as a result of necessary transformations. In addition, we assume that a set DK of formulas of the language $\mathcal{L}_{DK}$ and a set DtK of formulas of the language $\mathcal{L}_{Dt}$ are given. A workflow of data mining with association rules can be then described according to Fig. 2.

The first row in Fig. 2 means that an analytical question $\mathcal{Q}$ which can be solved by the procedure ASSOC is formulated using set DK of formulas of the language $\mathcal{L}_{DK}$. The set DtK of formulas of the language $\mathcal{L}_{DK}$ can also be used to formulate reasonable analytical questions.

A solution of $\mathcal{Q}$ starts with a definition of a set $\mathcal{S}(\Phi)$ of relevant association rules which have to be verified in $\mathcal{M}$, see row 2 in Fig. 2. The set $\mathcal{S}(\Phi)$ is given by a formula Φ of language $\mathcal{L}_{RAR}$. The formula Φ is a result of application of a procedure transforming formulas of $\mathcal{L}_{AQ}$ to formulas of $\mathcal{L}_{RAR}$.

Then, the procedure ASSOC is applied, see row 3 in Fig. 2. Experience with the procedure 4ft-Miner, which is an enhanced implementation of the ASSOC procedure, are given in [6, 7]. The application of ASSOC results into a set $True(\mathcal{S}(\Phi), \mathcal{M})$ of all association rules $\varphi \approx \psi$ which belong to $\mathcal{S}(\Phi)$ and which are true in $\mathcal{M}$.

```
 1   1 Formulate_Analytical_Question
 2     Define_Set_of_Relevant_Rules
 3   2 Apply ASSOC
 4     Apply CONCL
 5     IF Continue_ASSOC THEN
 6        BEGIN
 7        Modify Set_of_Relevant_Rules
 8        GOTO 2
 9        END
10     IF Continue_Analysis THEN GOTO 1
11     STOP
```

Figure 2. Association rule data mining workflow based on FOFRADAR

A next step is interpretation of the set $True(\mathcal{S}(\Phi), \mathcal{M})$. This is realized by the procedure $CONCL$, see row 4 in Fig. 2. Consequences of particular items of domain knowledge are used which means that the procedure $Cons$ is applied and several formulas of the language $\mathcal{L}_{Concl}$ are produced by the procedure $CONCL$.

One of formulas produced by $CONCL$ is a simple Boolean variable $Continue\_ASSOC$. If its value is setup as $true$, then the set $\mathcal{S}(\Phi)$ of relevant association rules which have to be verified is modified and the process of solution of the analytical question $\mathcal{Q}$ continues, ses rows 5 – 9 in Fig. 2. The modification of the set $\mathcal{S}(\Phi)$ is done by the procedure $ModifySet\_of\_Relevant\_Rules$ which uses experience from applications of the 4ft-Miner procedure.

If the value of $Continue\_ASSOC$ is setup to $false$, then the process of solution of the particular analytical question is terminated. Then a procedure $Continue\_Analysis$ is used to decide if an additional analytical question will be generated and solved.

There are first experiences with "manually driven" processes corresponding to procedures used in Fig. 2. The most important is experience with process corresponding to the $CONCL$ procedure, see [7]. We believe to get enough experience to run the whole data workflow process automatically as outlined in Fig. 2.

ACKNOWLEDGEMENTS

The work described here has been supported by Grant No. 201/08/0802 of the Czech Science Foundation.

REFERENCES

[1] R. Aggraval et al., *Fast Discovery of Association Rules*, 307–328, Advances in Knowledge Discovery and Data Mining, AAAI Press, Menlo Park, 1996.

[2] P. Hájek and T. Havránek, *Mechanising Hypothesis Formation - Mathematical Foundations for a General Theory*, Springer, Berlin, 1978

[3] M. Jalali-Heravi and R. Zaiane, *A study on interestingness measures for associative classifiers*, 1039–1046, Proceedings of the 2010 ACM Symposium on Applied Computing, ACM New York, 2010

[4] J. Rauch: Logic of association rules, *Applied Intelligence*, **22**, 9–28, 2005

[5] J. Rauch, *Consideration on a Formal Frame for Data Mining*, 562–569, Proceedings of Granular Computing 2011, IEEE Computer Society, Piscataway, 2011.

[6] J. Rauch and M. Šimůnek, *An Alternative Approach to Mining Association Rules*, 219–238, Data Mining: Foundations, Methods, and Applications, Springer-Verlag, Berlin, 2005.

[7] J. Rauch and M. Šimůnek, *Applying Domain Knowledge in Association-Rules Mining Process - First Experience*, 113–122, Foundations of Intelligent Systems, Springer-Verlag, Berlin, 2011.