Vysoká škola ekonomická v Praze **Fakulta informatiky a statistiky**
Katedra informačního a znalostního inženýrství **2009/2010**

ZADÁNÍ BAKALÁŘSKÉ PRÁCE

Autor práce: Bc. Viktor Nekvapil
Studijní program: Aplikovaná informatika
Obor: Informatika

Název tématu: **Aplikace procedury Ac4ft-Miner na medicínská data**

Rozsah práce: cca 30 - 40 stran

Zásady pro vypracování:

1. Seznámit se s procedurou Ac4FT-Miner
2. Definovat a vyřešit několik příkladů na zadaných datech a prezentovat pomocí systému SEWEBAR a/nebo ve formě dokumentů ve wordu
3. Shrnout zkušenosti z řešených příkladů a vypracovat stručnou metodiku aplikace procedury pro lékaře

Seznam odborné literatury:

1. RAUCH, Jan, ŠIMŮNEK, Milan. Action Rules and the GUHA Method: Preliminary Considerations and Results. Praha 14.09.2009 – 17.09.2009. In: Foundations of Intelligent Systems. Berlin : Springer Verlag, 2009, s. 76–87. ISBN 978-3-642-04124-2. ISSN 1867-8211.

Datum zadání bakalářské práce: **prosinec 2009**

Termín odevzdání bakalářské práce: **květen 2010**

Bc. Viktor Nekvapil
Řešitel

doc. RNDr. Jan Rauch, CSc.
Vedoucí práce

Ing. Vilém Sklenák, CSc.
Vedoucí ústavu

prof. Ing. Jan Seger, CSc.
Děkan FIS VŠE

VYSOKÁ ŠKOLA EKONOMICKÁ V PRAZE
FAKULTA INFORMATIKY A STATISTIKY

OBOR: INFORMATIKA

Aplikace procedury Ac4ft-Miner na medicínská data

BAKALÁŘSKÁ PRÁCE

Autor: Viktor Nekvapil
Vedoucí práce: doc. RNDr. Jan Rauch, CSc.
Jazyk práce: anglický

PRAHA 2010

PROHLÁŠENÍ

Prohlašuji, že jsem vypracoval samostatně bakalářskou práci na téma „Aplikace procedury Ac4ft-Miner na medicínská data“. Použitou literaturu a další podkladové materiály uvádím v přiloženém seznamu literatury.

V Praze dne 28. června 2010 .....

podpis studenta

PODĚKOVÁNÍ

Na tomto místě bych rád poděkoval vedoucímu mé bakalářské práce
doc. RNDr. Janu Rauchovi, CSc. za všestranné vedení, připomínky a cenné rady
při zpracování tématu a obzvláště pak za trpělivost při objasňování fungování procedury.

Abstrakt

Tato práce se zabývá dataminingovou procedurou Ac4ft-Miner implementovanou v systému LISp-Miner, který je vyvíjen na Katedře informačního a znalostního inženýrství Vysoké školy ekonomické v Praze. Cílem práce je jednak popsat proceduru jednoduchým, pro laiky (zejména lékaře) srozumitelným způsobem, dále pak aplikovat tuto proceduru na medicínská data a prezentovat příklady použití této procedury. Ze získaných zkušeností je pak cílem vytvořit metodologii použití procedury pro lékaře. Dosahování cílů je realizováno jednak použitím mnoha příkladů, které demonstrují teoretické koncepty na konkrétních datech, dále také snahou o jednoduchou vizualizaci úloh (analytických otázek) řešených procedurou.

Výsledkem práce je souhrnný, jednolitý text s řadou příkladů oddělených od souvislého textu, takže čtenář obeznámený s konkrétní tematikou může snadno pokračovat dále v textu, aniž by příklady musel projít. Dále je pak výsledkem práce návrh grafické podoby prezentace analytických otázek, která bude spolu s příklady využita dále v projektu SEWEBAR, jehož součástí je tato práce. Metodologie užití procedury pro lékaře má spíše podobu rad vhodných pro další výzkum této procedury, jelikož velká komplexnost procedury nedovolila formulaci obecnějších závěrů.

V první kapitole je popsán proces dolování dat z databází s využitím metodologie CRISP-DM, druhá kapitola prezentuje teoretické koncepty vztahující se k Ac4ft-Mineru, třetí kapitola objasňuje akční pravidla, jež jsou procedurou generována. Ve čtvrté kapitole jsou nastíněny možnosti zadávání úloh a interpretace výsledků procedury. Pátá kapitola obsahuje příklady zadání úloh, které byly řešeny procedurou, a také jsou prezentovány výstupy procedury. Šestá kapitola prezentuje metodologii užití procedury pro lékaře.

Abstract

This bachelor thesis deals with the data mining procedure Ac4ft-Miner, implemented in the LISp-Miner system, which is developed at the Department of Information and Knowledge Engineering at the University of Economics, Prague. The aim of this thesis is firstly to describe the procedure in a simple, understandable way. Secondly, the aim is to apply this procedure on the medical data and present examples of use of this procedure. Further aim is to create methodology of use for doctors from the experience obtained. The aims are reached by using a lot of examples, which demonstrate theoretical concepts on concrete data and by the pursuit of the simple visualisation of tasks (analytical questions) solved by the procedure.

The output of this thesis is a coherent text with lot of examples separated from the continuous text; so the reader familiar with a particular topic can skip the examples and proceed to the next issue. Further result of this thesis is an outline of the graphical presentation of analytical questions. Both the examples and the graphical presentation will be used further in the SEWEBAR project of which this thesis is one part. The methodology of use of the procedure for doctors is in the form of advices for use of the tool which should contribute to the further research which is needed. This is because of the high complexity of the procedure, which does not allow formulating general conclusions usable in the methodology.

Chapter 1 characterizes the overall process of Knowledge Discovery in Databases represented by the CRISP-DM Methodology. Chapter 2 presents theoretical concepts related to Ac4ft-Miner. Chapter 3 deals with action rules. Chapter 4 addresses possibilities of defining the input and interpretation of the output of the Ac4ft-Miner. Chapter 5 describes the research conducted on the real medical data set ADAMEK, states methodology and examples of the output. Chapter 6 summarises the experience obtained and formulates the methodology of use of Ac4ft-Miner for doctors.

Contents

Introduction

This bachelor thesis deals with the data mining procedure Ac4ft-Miner. Data mining is the analytical stage of knowledge discovery in databases (KDD). KDD is by [Fayadd et al., 1996] defined as *"the nontrivial extraction of implicit, previously unknown, and potentially useful information from data"*. According to [Fayyad et. al., 1996], KDD is an overall process consisting of data selection, pre-processing and transformation, data mining and the interpretation and evaluation of mined patterns which than lead to knowledge. KDD is interactive and iterative process, in other words, it is sometimes necessary to return to the previous step.

[Anand et al., 1996] enhances this rather technological view of other steps. He refers to KDD (as defined by Fayyad) as "*data mining*" because, as he claims, it reflects the industrial use of this term for the overall process. Firstly, after the managerial problem is identified, the human resources identification takes place. The three roles are as follows: Domain expert who has deep knowledge of the real problem; data mining expert who possesses knowledge of data mining methods; and data expert who has knowledge of data properties and data storaging. Next step is problem specification comprising of identification of task type and ultimate user of the knowledge. It is also important to gather all the data available, among others data from external resources such as statistical agencies. Moreover, the right data mining methodology has to be chosen, followed by data pre-processing and the knowledge discovery (data mining according to Fayyad's view). The final step, knowledge post-processing, encompasses the filtration of trivial and obsolete information and presentation of discovered knowledge in an appropriate form which is understandable for final users (e.g. owners of the data).

As mentioned above, data mining (in its narrower Fayyad's meaning) is the usage of analytical methods. It involves fitting models to observed data or determining patterns from it [Fayyad et al., 1996]. According to [Berka, 2003], the data mining has three main resources. The database technologies that enable to store large data sets; statistics that provides methods for analysing data; and, finally, machine learning, viewed e.g. as a searching through the area of potential solutions.

In this bachelor thesis, I will use the term "knowledge discovery in databases" (KDD) for the overall process described by [Anand et al., 1996] and the term "data mining" for applying methods to analysed data.

One of the analytical methods is the Ac4ft-Miner. It has three basic theoretical resources:

- the GUHA method
- association rules
- action rules

The GUHA method is method of exploratory analysis the aim of which is to provide all interesting facts derived from the analysed data. Association rules are the rules which express associations or correlation relationships among large set of data items. It shows attribute value conditions that appear frequently together in a given database [XLMiner, c2010]. Action rules express which action should be performed to improve the defined state (e.g. to improve the successfulness of a treatment).

Ac4ft-Miner can be thus described as follows:

Ac4ft-Miner finds rules that express which actions should be performed to improve the defined state. It achieves it by examining the dependencies among the data given as an input.

The Ac4ft-Miner is implemented in LISp-Miner System[1] developed at the Department of Information and Knowledge Engineering[2] at the University of Economics, Prague. The procedure is formally described in [Rauch, Šimůnek, 2009]; this thesis includes only an abbreviated description with the emphasis on user needs. The thesis is a part of the SEWEBAR[3] project "the goal of which is to study possibilities both of presenting results of data mining in the form of analytical reports and of disseminating resulting analytical reports through Semantic web" [SEWEBAR, c2010]. Note that all the information included in this thesis is related to the Ac4ft-Miner procedure unless stated differently.

The principal aims of the thesis are to:

1. provide an overview of the Ac4ft-Miner procedure
2. present examples of the use of the Ac4ft-Miner procedure on the real medical data set ADAMEK
3. convey the experience obtained
4. create a methodology of use of the Ac4ft-Miner for doctors

The target group of readers are especially doctors, i.e. people primarily not familiar with informatics and mathematics. That is why some of the facts have been simplified and many examples in boxes are provided to help the understanding of the issues. Those readers who are familiar with a particular topic can easily skip the boxes and proceed to the next issue.

The thesis is structured as follows:

Chapter 1 characterizes the overall process of KDD represented by the CRISP-DM Methodology. Examples illustrating the content of stages are included.

Chapter 2 introduces Boolean attributes, which are used in the Ac4ft-Miner, association rules and the GUHA method. Last section of the chapter integrates these three concepts into an implementation called 4ft-Miner, from which the Ac4ft-Miner is derived.

Third chapter discusses the issue of action rules. In the first section, the action rules introduced by [Ras, Wieczorkowska, 2000] are presented, second part comprises the G-action rules, which were inspired by Ras and which are implemented in the Ac4ft-Miner.

Chapter 4 mentions the possibilities of defining the input and interprets the output of the Ac4ft-Miner.

Chapter 5 describes the research carried out on the real medical data set ADAMEK, states the methodology and examples of the output.

Chapter 6 summarises the experience obtained and formulates the methodology of use of Ac4ft-Miner for doctors.

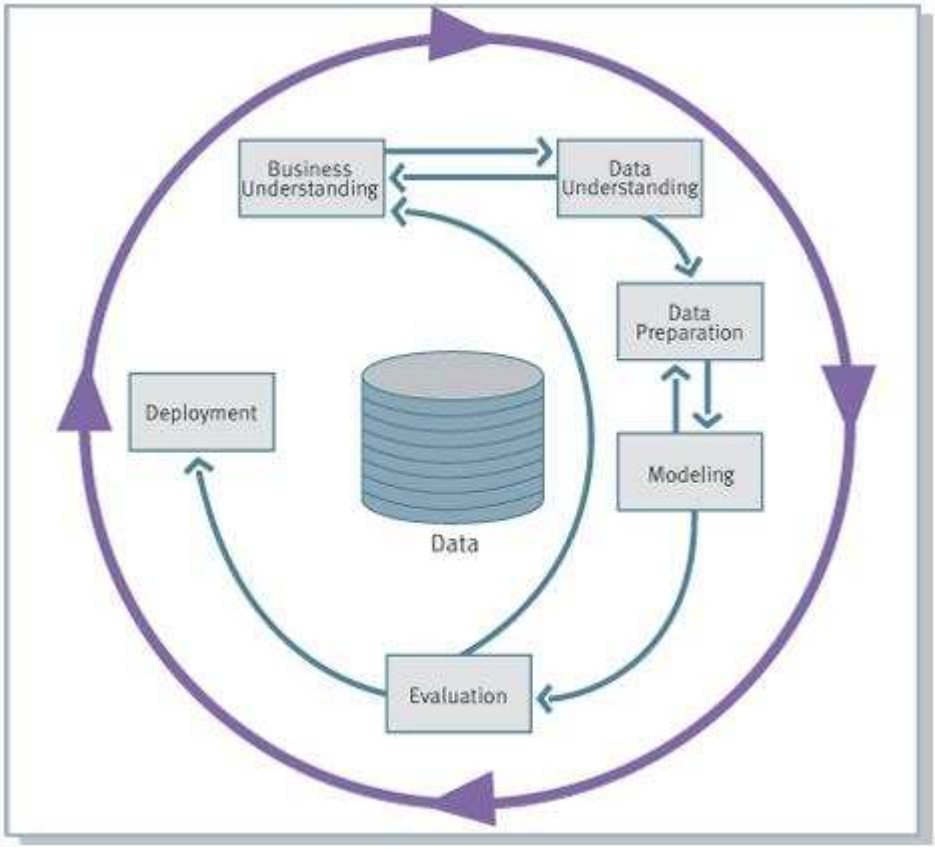
[1] You can find more information at http:// www.vse.cz/lispminer
[2] http:// kizi.vse.cz
[3] SEmantic WEB and Analytical Reports. You can get more information at http:// sewebar.vse.cz

1 The process of KDD - CRISP-DM Methodology

This chapter provides a deeper insight into the whole process of KDD, which was shortly mentioned in the introduction. I will demonstrate this process using the CRISP-DM Methodology [Chapman, et al., 2000].

CRISP-DM (**CR**oss **I**ndustry **S**tandard **P**rocess for **D**ata **M**ining) Methodology was created to address a need for standardisation in this field. Its aim was to facilitate the understanding of data mining as well as "demonstrate that data mining was sufficiently mature to be adopted as a key part of business processes" [Chapman, et al., 2000]. It has six phases (see Figure 1).

Figure 1: Phases of the CRISP-DM process



Source: [Chapman et al., 2000]

The arrows indicate that the order of the phases is not strictly given. It is often necessary to go back to the previous phase. The outer circle symbolizes the cyclical nature of the data mining in itself. After proceeding all the phases, it is necessary to employ the obtained experience in the next data mining task.

In the following sections, I will go through the phases and demonstrate their content on examples from a medical domain and the usage of the LISp-Miner system.

1.1 Business understanding

The aim of the first phase is to determine business objectives, i.e. to answer the question what the client (owner of the data) wants to accomplish by conducting data mining project. In this case, the term "business understanding" is perceived as gathering information in a specific area or domain; not only commercial-based projects are meant. It is also important to set success criteria; that is, when the outcome of the project is successful from a business point of view.

Furthermore, the first searching for available resources takes place. The resources include data resources, technical resources (e.g. computers, software), personnel resources (domain experts, data experts, data mining experts, technical support).

Other activities include risk assessment, costs and benefits and a creation of glossary of terminology, which helps with the communication between domain experts and "technical-oriented" experts.
It is also necessary to transform the business goals into data mining goals and to create project plan.
Some of the activities are shown in Box 1.

Box 1: Phase "Business understanding in a medical domain – selected activities

The primary "business" objective is to decrease the probability of having atherosclerosis. The data resources represent two sets of data collected in two different cities (small and larger one) in cardiological ambulances. Technical resources include standard technical equipment located at university and the LISp-Miner system, personnel resources include a doctor specialized in cardiology and a data expert and data mining experts from the university.
The data mining goal derived from the business objective can be as follows: to find rules which are connected to the risk factors of atherosclerosis.

1.2 Data understanding

In this phase,

- datasets are acquired from the data sources,
- data is described (e.g. format of the data, number of records with no value in each particular column ...),
- data is explored (finding of relationships among attributes, aggregations ...) and
- data quality is verified (number of errors in the dataset, missing values...)

An example of a data description is to be found in Appendix 1.

1.3 Data preparation

This phase includes

- data selection - selection of attributes (columns) as well as selection of records (rows) in a table, which are relevant to the data mining goals;

Box 2: Data selection

As a source of data, we can have sets collected by general practitioners. They include also data not relevant to our task – finding rules connected to risk factors of atherosclerosis (for example, the insurance company is not interesting for us). So this data will not be selected.

- data cleaning – errors, missing values and other issues reported in the previous phase are handled;

Box 3: Data preparation - cleaning

In LISp-Miner, there are three ways how to handle missing values [Berka, 2003]:

- Deleting – missing values are ignored, they are not involved in the algorithm
- Optimistic – missing values support the relationship we are interested in
- Secured – missing values do not support the relationship we are interested in

- data construction - derived attributes are created, some of the values are transformed to conform to the demands of a particular data mining tool;

Box 4: Data construction

An example of a derived attribute can be that of "Age", which is created by subtracting the year of birth of a patient from the current year.

In LISp-Miner, raw data from a database table are transformed into data matrix. Each column of a data matrix is created from a database column. The column in a data matrix is referred to as attribute and each value of an attribute is referred to as category. The usual definitions of attribute's categories are as follows:

- Each value – one category – the column is searched and each value is defined as one category (repeating values create only one category) – an example can be Type of therapy. Each possible value (e.g. diet, medicaments, operation, none) creates one category.
- Equidistant intervals – intervals of the same length are defined. An example of an appropriate attribute is Age. We can divide it into e.g. five-year intervals: Age ⟨0; 5), Age ⟨5; 10), Age ⟨10; 15) etc. The length of the interval can be various; the only limitation is the usefulness of such a category. It would not be useful to create only two intervals Age ⟨0; 60) and Age ⟨60; 120) because the results of the procedure would not make much sense (e.g. "The number of patients suffering from some disease in the Age ⟨0; 60) is higher than the number of patients in the Age ⟨60; 120)").
- Equifrequency intervals – intervals with approximately the same number of objects (same frequency) are created

For example, we can define attributes of objects (patients) and their categories as follows:

- Each value – one category: Sex (male), Sex (female)
- Age - equidistant intervals per 5 years; Age ⟨20;25), Age ⟨25;30), ..., Age ⟨60;65)
- Each value – one category: Type of therapy (diet), Type of therapy (medicaments), Type of therapy (operation), Type of therapy (none)
- Each value – one category: Success (yes), Success (no)
- Each value – one category: Genetic predisposition (yes), Genetic predisposition (no)
- other attributes (…)

- data integration – combining of data from different sources and tables, merging more tables to create one table

Box 5: Data integrated to one table

The integrated table comprising the data regarding a disease could look as in Figure 2.

Objects	Attributes						
Patient	Sex	Age	Type of therapy	Success	Genetic predisposition	City	…
1	male	42	none	no	no	Praha	…
2	female	61	diet	yes	no	Čáslav	…
3	female	24	operation	no	yes	Čáslav	…
4	male	54	medicaments	yes	no	Praha	…
…	…	…	…	…	…		…
632	female	57	medicaments	yes	no	Praha	…

Figure 2: Objects and their attributes

There are 632 patients in the table; each of them has been characterised in terms of his/her sex, age, type of therapy, whether this therapy was successful or not, whether he/she has genetic predispositions for this disease, a city where the examination took place and other attributes (...).

1.4 Modelling

This stage encompasses a selection of various modelling techniques (data mining algorithms such as Ac4ft-Miner) and their application on data. It is possible to select more than one modelling technique and compare their results. Specific modelling techniques require data in a specific form, so it often necessary to go back to the previous stage and modify the data. Before creating the model, we should test the quality and validity of a model on a training data. Then we have to adjust the parameters of a modelling tool and run it. The output of a tool is then evaluated in the terms of accuracy and generality of the model.

Examples of running a modelling tool (Ac4ft-Miner) are presented in chapters 5 and 6.

1.5 Evaluation

In this phase, we evaluate whether and to which degree the model meets the business objectives (success criteria) stated in the first phase. This phase includes also a decision, what to do next. If the model is sufficient, we can proceed to the next phase – deployment. If not, we have to return to previous stages. We can also decide to run another projects related to the current project.

1.6 Deployment

In this stage, the main goal is to apply the results of data mining to practical management. Therefore, it is important that the client (owner of the data) fully understands the results. The results can be either only presented in the form of analytical report or they can be utilised to create an automated system for classification of new cases. In the case of analytical reports, the results are successively used to improve the decision-making of an organisation.

Box 6: Deployment in a medical domain

Once the analytical report is presented to the doctors, they can see which factors influence the risk factors of atherosclerosis. They can use these results in the ambulance of preventive cardiology to decrease the probability of their patients having manifestations of atherosclerosis.
The results can also be used to create a software available via the internet to classify patients into groups according to their risk of having manifestation of atherosclerosis. This can help reveal risk patients without the need to visit doctor.

2 Prerequisites

In this chapter, theoretical concepts of Ac4ft-Miner are presented. In the first section, the Boolean attributes are clarified. The second section deals with association rules, the third one presents the GUHA method and the fourth section introduces an approach of combining association rules and the GUHA method.

2.1 Boolean attributes

LISp-Miner works with the data in the form of basic Boolean attributes and derived Boolean attributes. Each basic Boolean attribute is created from an attribute plus one or more of its categories. An example of a Boolean attribute is a *type of therapy being diet or medicaments* (symbolic expression: Type of therapy (diet, medicaments)). Set of one or more categories of a particular basic Boolean attribute is also called a coefficient of a basic Boolean attribute (the coefficient of the basic Boolean attribute Type of therapy from previous example is the set {diet, medicaments}). We can say about each Boolean attribute whether it is true or not for each object in a data matrix.

Using propositional connectives ∨, ∧ and ¬, derived Boolean attributes are created from basic Boolean attributes. Example is shown in Box 7.

Box 7: Basic Boolean attributes and derived Boolean attributes

Examples of basic Boolean attributes and derived Boolean attributes which could be created from the table given in Box 5 are shown in Figure 3.

Objects	Basic Boolean attributes		Derived Boolean attributes		
Patient	Sex (male)	Type of therapy (operation)	Sex (male) ∨ success (yes)	Genetic predisp. (no) ∧ Age ⟨50,60)	Sex (male) ∧ (Type of therapy (diet) ∨ Type of therapy (medicaments)) ∧ ¬Age ⟨50,60)
1	true	false	true	false	false
2	true	false	true	false	true
3	false	true	false	false	false
4	true	false	true	true	false
…	…	…	…	…	…
632	false	false	true	true	false

Figure 3: Boolean attributes and derived Boolean attributes

propositional connectives:
∨ logical disjunction, OR
∧ logical conjunction, AND
¬ negation, NOT

2.2 Association rules

Association rules are rules which express associations or correlation relationships among the large set of data items. It shows attribute value conditions that appear frequently together in a given database. They were largely popularised by Agrawal (e.g. in [Agrawal et al., 1993]). He used association rules in an analysis of shopping baskets, where he analysed customers´ transactions in a supermarket. The association rule is an implication

A => B, where A (called antecedent) and B (called consequent) are sets of items.

This rule can be understood as follows: In a particular data set, if a transaction includes items of A, it often includes also items of B.
To measure the strength of the rule, c*onfidence* is introduced:

$$confidence = \frac{number\ of\ transactions\ containing\ both\ A\ and\ B}{number\ of\ transactions\ containing\ A}$$

Confidence of e.g. 0.6 expresses that 60 % of transactions containing A also contain B.

To measure statistical significance, *support* is defined as

$$support = \frac{number\ of\ transactions\ containing\ both\ A\ and\ B}{number\ of\ all\ transactions\ in\ the\ database}$$

If support is too low, this means the rule is not worth considering, because there are too few transactions in the database in comparison to the size of the database (transactions do not support the rule).

Box 8: Example of an association rule

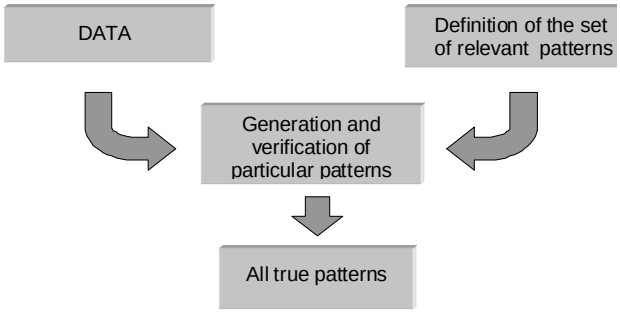
Beer, ham => sugar confidence = 0.6, support = 0.2

Interpretation:
60 % of customers buying beer and ham also buy sugar. Transactions containing both beer and ham comprise 20 % of all transactions in a database and this rule is thus for us worth considering.

2.3 GUHA method

The goal of the GUHA method is to provide all interesting facts available in the analysed data. The GUHA method is realised by GUHA procedures (see Figure 4). The input of GUHA procedures consists of analysed data and a simple definition of a set of relevant (i.e. potentially interesting) patterns. The procedure generates each particular pattern and tests whether it is true in the analysed data. The output of this procedure is a set of patterns which are true in the analysed data and are prime (they do not follow from other more simple rules)[4] [Rauch, Šimůnek, 2009].

Figure 4: The GUHA procedure



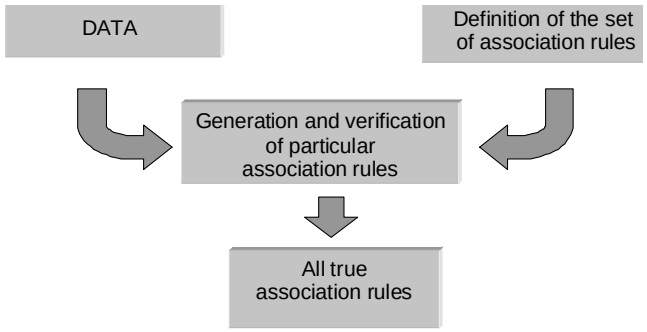
Source: [Rauch, Šimůnek, c2010]

[4] Example on prime pattern in chapter 4.2

2.4 GUHA method and Association rules

When we implement the GUHA method to generate association rules (which do so called ASSOC procedures, e.g. the 4ft-Miner), the whole process looks as depicted in Figure 5. The association rules mined by these procedures are, however, more general than the "classical" association rules introduced in chapter 2.2.

Figure 5: GUHA procedure ASSOC



Source: [Rauch, Šimůnek, c2010]

Association rules that are mined by the ASSOC procedures are the rules in the form

$$\varphi \approx \psi,$$

where both φ and ψ are Boolean attributes or derived Boolean attributes. φ is called antecedent, ψ is called succedent. We can understand the relationship between antecedent and succedent as an association. They are associated in a way given by the 4ft-quantifier ≈. 4ft-quantifier is mathematical expression of the dependency of φ and ψ. It is a condition that enables us to say whether the relationship between φ and ψ is true or not. All quantifiers use the four-fold table depicted in Figure 6.

	ψ	¬ ψ
φ	a	b
¬ φ	c	d

Figure 6: Four-fold table

As φ and ψ, Boolean attributes are used (see chapter 2.1); a, b, c and d are natural numbers.

- a is a number of objects which satisfy both φ and ψ;
- b is a number of objects which satisfy φ but not ψ;
- c is a number of objects not satisfying φ but satisfying ψ;
- d is a number of objects not satisfying both φ and ψ. An example is shown in Box 9.

The dependency between φ and ψ is expressed by 4ft-quantifiers. As a simple example we use quantifier $\Rightarrow_{p,B}$ that is defined by the condition

$$\frac{a}{a+b} \geq p \wedge a \geq B, \qquad \text{where } 0 < p \leq 1 \text{ and } B > 0.$$

The association rule

$$\varphi \Rightarrow_{p,B} \psi$$

means that at least 100p objects satisfying φ also satisfy ψ ($\frac{a}{a+b} \geq$ p in the four-fold contingency table, also referred to as *confidence*), and there are at least B objects satisfying both φ and ψ ("a" in the four-fold table). Example of a four-fold table and a usage of a quantifier is presented in Box 9.

There are many other 4ft-quantifiers defined in the LISp-Miner System. The procedure that mines for these association rules is called 4ft-Miner and is also implemented in LISp-Miner System. Ac4ft-Miner also uses association rules but in a way described in the next chapter.

Box 9: Example of a four-fold table and the usage of a quantifier

If we go back to Box 7, there we find derived Boolean attribute Genetic predisp. (no) ∧ Age ⟨50; 60), and basic Boolean attribute Type of therapy (operation). Let us make a four-fold contingency table for those two attributes (Figure 7). It is important to note that not only "visible" rows (patients) from Box 7 were included in this table.

	Type of therapy (operation)	¬ Type of therapy (operation)
Genetic predisp. (no) ∧ Age ⟨50; 60)	84	65
¬ (Genetic predisp. (no) ∧ Age ⟨50; 60))	205	187

Figure 7: Four-fold table

Now we have to express the dependency which should have both attributes. We use the *founded implication* described above. If we substitute numbers from our four-fold table, we have

$\frac{84}{84+65} \geq p \wedge 84 \geq$ B. The p and B have to be defined prior to the generation of the rules by the user and they indicate what strength the user wants the rules to have. Let us assume that we define p=0.7 (which means that the condition has to be satisfied by at least 70% of patients) and B=80 (the condition is true for at least 80 patients).

For the four-fold table in Figure 7, the calculated p = $\frac{84}{84+65} = 0.56$. So the condition is not fulfilled (our defined p = 0.7 is higher than the calculated p = 0.56) and thus the rule is not true in the data matrix (this meaning that it is not true that patients between the age of 50 and 60 with no genetic predispositions have their type of therapy operation). The rule would not be a part of the output, because the output contains only rules which are true in the data matrix.

But if we define p=0.5, the condition is satisfied and we can say that it is true in the data matrix that patients between the age of 50 and 60 with no genetic predispositions do have their type of therapy operation.

A symbolic expression of this rule is

Genetic predisp. (no) ∧ Age ⟨50; 60) $\Rightarrow_{0.56,84}$ Type of therapy (operation)

This rule can be also understood in a way that there are 84 patients in a database who represent 56 % of all patients in the age between 50 and 60 with no genetic predisposition, who have their type of therapy operation.

3 Action rules

Action rules express which action should be performed to improve the defined state. Action rules were firstly proposed in [Ras, Wieczorkowska, 2000], chapter 3.1 presents this approach. In chapter 3.2, the implementation of action rules in the GUHA method is introduced. The procedure is called Ac4ft-Miner and it mines for the so-called G-action rules.

3.1 Action rules according to [Ras, Wieczorkowska, 2000]

Action rules suggest a change in behaviour that can bring us an advantage. They assume two sets of attributes in the database – stable attributes and flexible attributes. Flexible attributes are the attributes which can be somehow changed by the user (that is, it is possible to change the behaviour in reality and in this way to change the attribute), stable attributes cannot be changed. An example of a stable attribute in a medical database is Sex; an example of a flexible attribute is Type of therapy. Furthermore, there is an attribute called decision which is a result of a rule.

A pattern of an association rule can be expressed as follows [Ras,2007]:

$$R: (A_1 = \omega_1) \wedge \ldots \wedge (A_Q = \omega_Q) \wedge (B_1, \alpha_1 \rightarrow \beta_1) \wedge \ldots \wedge (B_P, \alpha_P \rightarrow \beta_P) \Rightarrow (D, k_1 \rightarrow k_2)$$

- Where $(A_1, \ldots, A_Q)$ are stable attributes
- $(\omega_1,\ldots,\omega_Q)$ are values of stable attributes $(A_1, \ldots, A_Q)$
- $\{B_1, \ldots, B_P\}$ are flexible attributes
- $\{(\alpha_1 \rightarrow \beta_1), \ldots, (\alpha_P \rightarrow \beta_P)\}$ are changes of values of flexible attributes $\{B_1, \ldots, B_P\}$
- D is a decision
- $(k_1 \rightarrow k_2)$ is a change of decision from k_1 to k_2

Analogously to association rules, support and confidence are defined. Assume

- CPL(R)...number of objects matching $(\omega_1,\ldots,\omega_Q,\alpha_1,\ldots,\alpha_P, k_1)$, i.e. number of objects matching the state before a change which also match the state of decision before the change
- CPR(R): number of objects matching $(\omega_1,\ldots,\omega_Q,\beta_1,\ldots, \beta_P, k_2)$, i.e. number of objects matching the state after a change which also match the state of decision after the change
- CVL(R): number of objects matching $(\omega_1,\ldots,\omega_Q,\alpha_1,\ldots,\alpha_P)$, i.e. number of all objects matching the state before a change
- CVR(R): number of objects matching $(\omega_1,\ldots,\omega_Q,\beta_1,\ldots, \beta_P)$, i.e. number of all objects matching the state after a change
- $\text{LeftSup(R)} = \frac{\text{CPL(R)}}{n}$, where n is a total number of objects in the database
- $\text{RightSup(R)} = \frac{\text{CPR(R)}}{n}$

then support is defined as $\text{Sup(R)} = \text{LeftSup(R)} = \frac{\text{CPL(R)}}{n}$,

and confidence is defined as $\text{Conf(R)} = \frac{\text{CPL(R)}}{\text{CVL(R)}} * \frac{\text{CPR(R)}}{\text{CVR(R)}}$.

An example is provided in Box 10.

Box 10: Example of an action rule

Sex(female) $\wedge$ Age $\langle 50,60)$ $\wedge$ Type of therapy (diet $\rightarrow$ medicaments) $\Rightarrow$ Success (no $\rightarrow$ yes)

Word expression of this example:

If somebody is a woman, her age is between 50 and 60 and the therapy is changed from diet to medicaments, the success of a treatment changes from unsuccessful to successful.

Stable attributes in this example are Sex and Age, flexible attribute is Type of therapy and decision is Success.

Assume

- n = 632 There are 632 patients in a database
- CPL(R) = 42 Number of objects matching (female, (50; 60), diet, no)
- CPR(R) = 74 Number of objects matching (female, (50; 60), medicaments, yes)
- CVL(R) = 90 Number of objects matching (female, (50; 60), diet)
- CVR(R) = 105 Number of objects matching (female, (50; 60), medicaments)

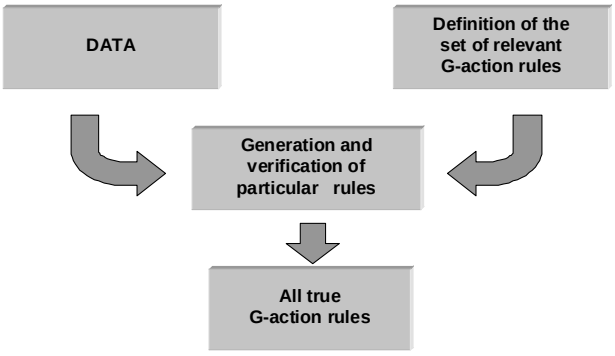
then

- Sup(R) = LeftSup(R) = $\frac{\mathrm{CPL(R)}}{n} = \frac{42}{632} = 0.07$
- Conf(R) = $\frac{\mathrm{CPL(R)}}{\mathrm{CVL(R)}} * \frac{\mathrm{CPR(R)}}{\mathrm{CVR(R)}} = \frac{42}{90} * \frac{74}{105} = 0.33$

3.2 G – action rules

G-action rules were inspired by action rules introduced in the previous chapter (3.1).They also assume stable and flexible attributes. G-action rules are the rules suggesting an action which are mined using the GUHA method. The procedure that mines for G-action rules is called Ac4ft-Miner. The input of Ac4ft-Miner is similar to the 4ft-Miner: a data matrix and a definition of the set of relevant G-action rules from which true rules will be selected (see Figure 8). The simplified definition of the set of relevant G-action rules is presented in Box 11.

Figure 8: GUHA procedure Ac4ft-Miner



Source: [Rauch, Šimůnek, c2010]

Box 11: Simplified definition of the set of relevant G-action rules and simplified output of the Ac4ft-Miner

Let us assume we are interested in the question which changes of type of therapy influence the success of a treatment. We will be also interested in the sex of the patient and their age.

A simplified definition is as follows:
Find all true G-action rules from the data matrix concerning stable attributes Sex and Age and flexible attributes Type of therapy and Success.

Then the output could be as follows:

If somebody is a woman, her age is between 50 and 60 and the therapy is changed from diet to medicaments, the success of a treatment changes from unsuccessful to successful.

If somebody is over 60 and the therapy is changed from operation to diet, the success of a treatment changes from successful to unsuccessful.

If somebody is a man and the therapy is changed from diet or medicaments to operation, the success of a treatment changes from unsuccessful to successful.
…

G-action rule is the rule in a form

$$\varphi_{St} \wedge \Phi_{Chg} \approx^* \psi_{St} \wedge \Psi_{Chg}$$

- where φ_{St} is Boolean attribute called stable antecedent (or antecedent stable part),
- Φ_{Chg} is called change of antecedent (or antecedent flexible part),
- ψ_{St} is Boolean attribute called stable succedent (or succedent stable part),
- Ψ_{Chg} is called change of succedent (or succedent flexible part),
- $\approx^*$ is called Ac4ft-quantifier (the analogy to the 4ft-quantifier described above).

Both the change of antecedent Φ_{Chg} and the change of succedent Ψ_{Chg} are changes of coefficient of Boolean attributes (e.g. from Type of therapy (diet) to Type of therapy (medicaments)). The attributes in Φ_{Chg} are called independent attributes and the attributes in Ψ_{Chg} are called dependent attributes. The action rule expresses what happens with objects satisfying stable conditions φ_{St} and ψ_{St} when we change their flexible independent attributes in a way given by Φ_{Chg}. The initial state of Φ_{Chg} is characterised by Boolean attribute I (Φ_{Chg}), the final state is characterised by Boolean attribute $F(\Phi_{Chg})$. Similarly, the initial state of dependent attribute Ψ_{Chg} is characterised by Boolean attribute $I(\Psi_{Chg})$, the final state is characterised by Boolean attribute $F(\Psi_{Chg})$.

The effect is described by two association rules:

R_I: $\varphi_{St} \wedge I(\Phi_{Chg}) \approx_I \psi_{St} \wedge I(\Psi_{Chg})$ $\quad$ R_F: $\varphi_{St} \wedge F(\Phi_{Chg}) \approx_F \psi_{St} \wedge F(\Psi_{Chg})$

We can denote $\varphi_{St} \wedge I(\Phi_{Chg})$ as φ_I, $\psi_{St} \wedge I(\Psi_{Chg})$ as ψ_I, $\varphi_{St} \wedge F(\Phi_{Chg})$ as φ_F and $\psi_{St} \wedge F(\Psi_{Chg})$ as ψ_F. The rules are now simplified as follows:

R_I: $\varphi_I \approx_I \psi_I$ $\quad$ R_F: $\varphi_F \approx_F \psi_F$

The first rule R_I describes the initial state; the second rule R_F describes the final state induced by the change of independent flexible attribute Φ_{Chg}. The initial relation between antecedent (φ_I) and

succedent (ψ_I) is described by the 4ft-quantifier $\approx_I$, the final relation between antecedent (φ_F) and succedent (ψ_F) is described by the 4ft-quantifier $\approx_F$.

Both the rules are evaluated using their own four-fold table, see Figure 9.

	ψ_I	$\neg\ \psi_I$
φ_I	a_I	b_I
$\neg\ \varphi_I$	c_I	d_I

	ψ_F	$\neg\ \psi_F$
φ_F	a_F	b_F
$\neg\ \varphi_F$	c_F	d_F

Figure 9: Four-fold tables for $\mathbf{R}_I$ and $\mathbf{R}_F$

The truthfulness of the whole action rule $\varphi_{\text{St}} \wedge \Phi_{\text{Chg}} \approx^* \psi_{\text{St}} \wedge \Psi_{\text{Chg}}$ is defined by Ac4ft-quantifiers. There are several of them. As an example, we can use Ac4ft-quantifier $\Rightarrow^{F>I}_{q,B_1,B_2}$ which is defined by the condition

$$\frac{a_F}{a_F+b_F} - \frac{a_I}{a_I+b_I} \geq \boldsymbol{q} \wedge a_I \geq B_I \wedge a_F \geq B_F, \text{ where } 0 < q \leq 1, B_I > 0 \text{ and } B_F > 0.$$

- B_I is desired number of objects (rows in data matrix) satisfying both the initial antecedent and initial succedent,
- B_F is desired number of objects satisfying both the final antecedent and final succedent.
- q is desired percentage difference between the initial and final state. It says by how many percentage points the confidence of the final rule is higher in comparison with the initial rule.

If this Ac4ft-quantifier is true in a data matrix, its effect can be described by two association rules:

R_I: $\varphi_I \Rightarrow_{=,\,p,\,B} \psi_I$ $\qquad$ R_F: $\varphi_F \Rightarrow_{=,\,p+\mathbf{q},\,B} \psi_F$

- In R_I, the 4ft-quantifier $\Rightarrow_{=,\,p,\,B}$ is defined by the condition $\frac{a}{a+b} = p \wedge a = B$,
- In R_F, the 4ft-quantifier $\Rightarrow_{=,\,p+\mathbf{q},\,B}$ is defined by the condition $\frac{a}{a+b} = p + q \wedge a = B$.

Each Ac4ft-quantifier has two 4ft-quantifiers which describe the effect of a rule. There is not a complete list of Ac4ft-quantifiers as they can be created from various 4ft-quantifiers. Ac4ft-quantifiers are the same as SD4ft-quantifiers which are described in [Kodym, 2007].

The Ac4ft quantifier $\Rightarrow^{F>I}_{q,B_1,B_2}$ is suitable when we have "positive" attribute (or attributes) in succedent. By "positive" it is meant that it is desirable to increase probability (truthfulness) of this attribute. An example of this attribute is the attribute Success (yes).

When it is desirable to decrease the probability of a succedent (it is "negative"), the Ac4ft quantifier $\Rightarrow^{I>F}_{q,B_1,B_2}$ is suitable. It is defined by the condition

$$\frac{a_I}{a_I+b_I} - \frac{a_F}{a_F+b_F} \geq q \wedge a_I \geq B_I \wedge a_F \geq B_F, \text{ where } 0 < q \leq 1, B_I > 0 \text{ and } B_F > 0.$$

An example of a "negative" attribute can be Success (no).

I define a new Ac4ft quantifier $\Rightarrow^{I \lesseqgtr F}_{q,B_1,B_2}$, which uses either of the quantifiers defined above. The relationship between antecedent and succedent must satisfy either the condition defined by Ac4ft-quantifier $\Rightarrow^{F > I}_{q,B_1,B_2}$ or the condition defined by $\Rightarrow^{I > F}_{q,B_1,B_2}$. The Ac4ft quantifier $\Rightarrow^{I \lesseqgtr F}_{q,B_1,B_2}$ is thus defined by the condition:

$$\left| \frac{a_I}{a_I + b_I} - \frac{a_F}{a_F + b_F} \right| \geq q \wedge a_I \geq B_I \wedge a_F \geq B_F, \qquad \text{where } 0 < q \leq 1, B_I > 0 \text{ and } B_F > 0.$$

I will use this quantifier in my case study.

Both the generation and verification of action rules is automated, the task of a user is only to insert the restrictive conditions in order not to generate and verify all the rules that are possible to be created in a data matrix (which would be indeed time and source consuming). The input of Ac4ft Miner procedure is introduced in the next chapter.

Box 12: Example of definition of the set of relevant action rules and verification of a rule

Firstly, we have to define a set of relevant action rules from which particular rules will be created. (a simplified example was presented in Box 11, in this place the more precise definition is given). We define the stable antecedent as a combination of attributes Sex and Age, the flexible antecedent as Type of therapy and the stable succedent as Success. We do not want to use the flexible succedent. As Ac4ft-quantifier we want to use $\Rightarrow^{F > I}_{q,B_1,B_2}$, i.e. that the confidence of the final rule must be higher than the confidence of the inital rule. We want at least 15 patients to satisfy both the initial antecedent and the initial succedent ($B_I = 15$) and at least 15 patients to satisfy both the final antecedent and the final succedent ($B_F = 15$) and the absolute difference of confidence of the initial and the final rule to be at least 30 percentage points (q = 0.3).

An output of such a definition can consist of several action rules. One of the rules can be as follows:

Sex (female) $\wedge$ Age $\langle 50; 60)$ $\wedge$ Type of therapy (diet $\rightarrow$ medicaments) $\Rightarrow_{0.386,17,25}$ Success (yes)

The stable antecedent of this rule is the expression Sex (female) $\wedge$ Age $\langle 50; 60)$, the flexible antecedent (independent attribute) is the Boolean attribute Type of therapy, which is changed from the initial state Type of therapy (diet) to the final state Type of therapy (medicaments). The stable succedent is Success (yes), and the succedent flexible part is missing in this example. The two association rules which describe the effect of action rule are as follows:

R_I: Sex (female) $\wedge$ Age $\langle 50; 60)$ $\wedge$ Type of therapy (diet) $\Rightarrow_{=,\ 0.607,17}$ Success (yes)

R_F: Sex (female) $\wedge$ Age $\langle 50; 60)$ $\wedge$ Type of therapy (medicaments) $\Rightarrow_{=,\ 0.893,25}$ Success (yes)

Continues on the next page

Continuation from the previous page

These rules were obtained from four-fold table for R_I (see Figure 10) and from four-fold table for R_F (Figure 11). The tables are built in the same way as in Figure 7.

	Success (yes)	¬ Success (yes)
Sex (female) ∧ Age ⟨50; 60) ∧ Type of therapy (diet)	17	11
¬ (Sex (female) ∧ Age ⟨50; 60) ∧ Type of therapy (diet))	205	398

Figure 10: Four-fold table for R_I

	Success (yes)	¬ Success (yes)
Sex (female) ∧ Age ⟨50; 60) ∧ Type of therapy (medicaments)	25	3
¬ (Sex (female) ∧ Age ⟨50; 60) ∧ Type of therapy (medicaments))	267	382

Figure 11: Four-fold table for R_F

In case of the rule R, the condition is

$$\frac{25}{25+3} - \frac{17}{17+11} \geq 0.3 \wedge 17 \geq 15 \wedge 25 \geq 15.$$

In case of R_I, the 4ft-quantifier $\Rightarrow_{=,\ p,\ B}$ created from the four-fold table (Figure 10) is $\frac{17}{17+11} = p = 0.607$.

In case of R_F, it is $\frac{25}{25+3} = p+q = 0.893$ and thus $q = 0.893 - 0.607 = 0.386$.

As we can see, the action rule R meets all the conditions set and is thus true in a data matrix and thus appears in the output of the Ac4ft-Miner procedure. The meaning of the whole rule R can be expressed as follows:

If the therapy is changed from diet to medicaments among the female patients between the age of 50 and 60, the probability of a successful treatment increases by 38.6 percentage points.

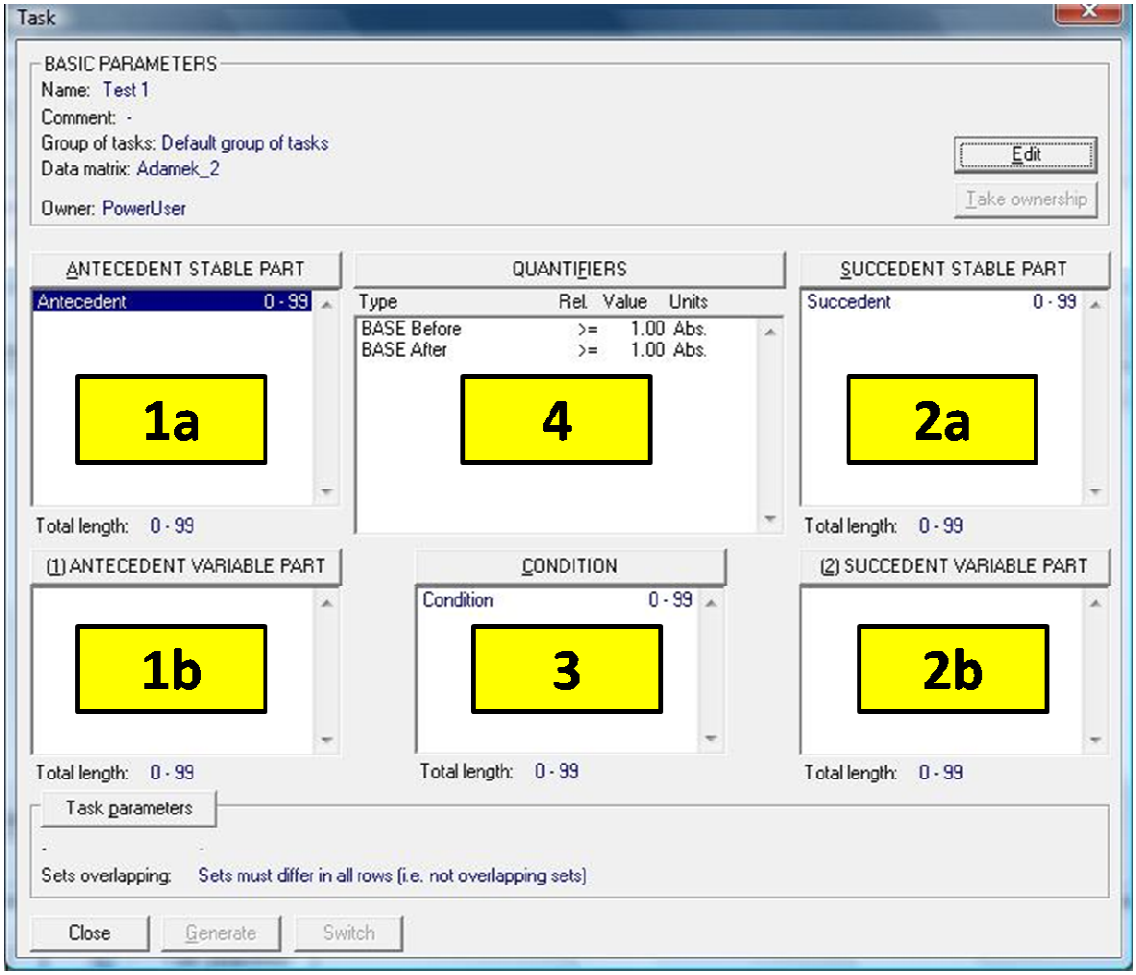
4 Input and output in Ac4ft-Miner

In this chapter I summarise and elaborate the information from previous chapters concerning the input and output in Ac4ft-Miner. It is a prerequisite for understanding of the formulation of analytical questions in the next chapter. This chapter is based on [Rauch, 2005].

4.1 Input

The input consists of following parts:

- Data matrix
- Entering a set of relevant rules which should be generated and verified (see Figure 12)
    1. Entering a set of relevant antecedents
        a. Antecedent stable part
        b. Antecedent variable part
    2. Entering a set of relevant succedents
        a. Succedent stable part
        b. Succedent variable part
    3. Entering a set of relevant conditions
    4. Entering an Ac4ft-quantifier



4.1.1 Data matrix

As a data matrix, any relational database can be attached to the LISpMiner via the ODBC interface. An example of such a database system is Microsoft Access.

4.1.2 Entering a set of relevant rules

4.1.2.1 Entering a set of relevant antecedents, succedents and conditions

Antecedent, succedent and condition are called *cedents*. The cedents can be divided into *partial cedents*. The partial cedent is a conjunction or disjunction of *literals.* The literal is a basic Boolean attribute (i.e. positive literal) or a negation of basic Boolean attributes (i.e. negative literal).

Condition enables us to set additional restrictions which every column of a data matrix (each object) must fulfil. It means that only the object which matches the condition is allowed to appear in a four-fold table of antecedent and succedent. For example, the condition Sex (male) restricts the data matrix only to men; all objects having in column Sex other values than male are ignored.

Definition of the relevant set of literals

For each attribute, we should specify which set of literals will be created. This definition is determined by:

1. Minimal and maximal length of a literal.
2. The type of coefficient – subsets, intervals, cyclical intervals, left cuts, right cuts, cuts, one particular value
3. One from the following options:
   - Generate only positive literals – no literals with negation are created
   - Generate only negative literals – only literals with negation are created
   - Generate both positive and negative literals

1. The length of a literal is defined as the number of categories that one literal has (for example, Type of therapy (diet, medicals) is a literal of the length 2).

2.Types of coefficients

Subsets. Creation of all possible combinations of categories of the defined length (the order does not matter)

Example: Create literals of the attribute Type of therapy with its categories {diet, medicaments, operation, none} with minimal length 1 and maximal length 2:

Type of therapy (diet), Type of therapy (medicaments), Type of therapy (operation), Type of therapy (none), Type of therapy (diet, medicaments), Type of therapy (diet, operation), Type of therapy (diet, none), Type of therapy (medicaments, operation), Type of therapy (medicaments, none), Type of therapy (operation, none).

Intervals. In this case, sequences of the defined length are created.

Example: Create literals of the attribute Age with its categories {⟨20; 30), ⟨30; 40), ⟨40; 50), ⟨50; 60), ⟨60; 70)} with minimal length 2 and maximal length 3.

Age [⟨20; 30), ⟨30; 40)], Age [⟨30; 40), ⟨40; 50)], Age [⟨40; 50), ⟨50; 60)], Age [⟨50; 60), ⟨60; 70)],
Age [⟨20; 30), ⟨30; 40), ⟨40; 50)], Age [⟨30; 40), ⟨40; 50), ⟨50; 60)], Age [⟨40; 50), ⟨50; 60), ⟨60; 70)].

Note: We can simplify the expression Age [⟨40; 50), ⟨50; 60), ⟨60; 70)] to Age ⟨40; 70)

Cyclical intervals. Sequences of the defined length are created, cycles are permitted.

Example: Create literals of the attribute Day with its categories {sun, mo, tue, we, thu, fri, sat} with minimal length 3 and maximal length 4.

Day (sun, mo, tue), Day (mo, tue, we), Day (tue, we, thu), Day (we, thu, fri), Day (thu, fri, sat), Day (fri, sat, sun), Day (sat, sun, mo),
Day (sun, mo, tue, we), Day (mo, tue, we, thu), Day (tue, we, thu, fri), Day (we, thu, fri, sat), Day (thu, fri, sat, sun), Day (fri, sat, sun, mo), day (sat, sun, mo, tue).

Left cuts. Sequences containing only the first category are created.

Example: Create literals of the attribute Age with its categories {⟨20; 30), ⟨30; 40), ⟨40; 50), ⟨50; 60), ⟨60; 70)} with maximal length 4 (minimal length is by default 1).

Age ⟨20; 30), Age [⟨20; 30), ⟨30; 40)], Age [{⟨20; 30), ⟨30; 40), ⟨40; 50)],
Age [⟨20; 30), ⟨30; 40), ⟨40; 50), ⟨50; 60)].

Right cuts. Sequences containing only the last category are created.

Example: Create literals of the attribute Age with its categories {⟨20; 30), ⟨30; 40), ⟨40; 50), ⟨50; 60), ⟨60; 70)} with maximal length 4 (minimal length is by default 1).

Age ⟨60; 70), Age [⟨60; 70), ⟨50; 60)], Age [⟨60; 70), ⟨50; 60), ⟨40; 50)],
Age [⟨60; 70), ⟨50; 60), ⟨40; 50), ⟨30; 40)].

Cuts. Generation of both left cuts and right cuts.

One particular value. Only one literal with a particular category will be used.

Example: Create literal of the attribute Type of therapy with its category {diet, medicaments, operation, none} containing only operation.

Type of therapy (operation).

4.1.2.2 Entering an Ac4ft-quantifier

There are various options how to enter an Ac4ft-quantifier. In my case study I will only use the following quantifiers:

- BASE Before – number of objects satisfying both antecedent and succedent in the initial rule
- BASE After – number of objects satisfying both antecedent and succedent in the final rule
- Founded Implication – Difference of quantifier values – this quantifier is characterised by the relationship $\left| \frac{a_I}{a_I + b_I} - \frac{a_F}{a_F + b_F} \right| \geq q$, where q is the difference of the values of quantifiers (see chapter 3.2).

4.2 Output

The output of the Ac4ft-Miner consists of a set of prime action rules. The rule is prime if it is true in the analysed data and does not follow from other more simple action rules. See Box 13 for an example.

Box 13: Prime action rule

The rule

Sex (female) $\wedge$ Age [$\langle$40; 50), $\langle$50; 60)] $\wedge$ Type of therapy (diet $\rightarrow$ medicaments) $\Rightarrow_{0.386,17,25}$ Success (yes)

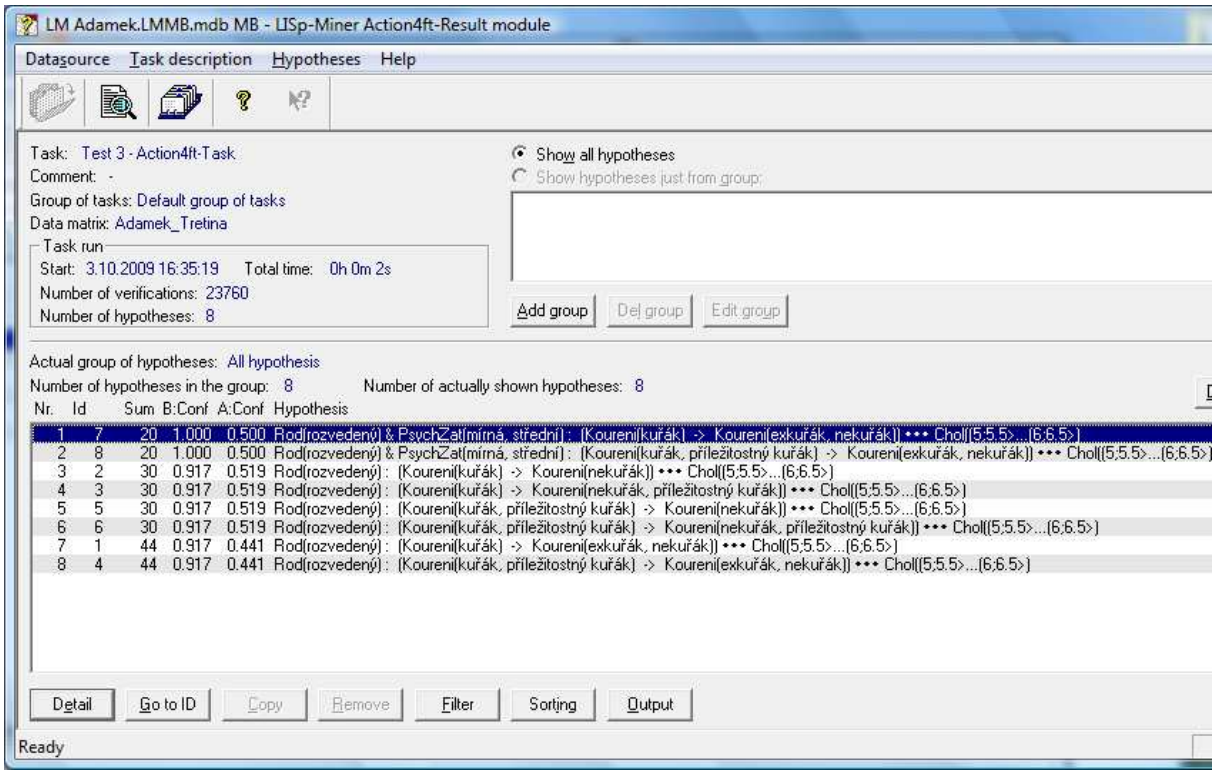
is not prime because it follows from the rule

Sex (female) $\wedge$ Age $\langle$50; 60) $\wedge$ Type of therapy (diet $\rightarrow$ medicaments) $\Rightarrow_{0.386,17,25}$ Success (yes).

Note that the attribute in the first rule Age [$\langle$40; 50), $\langle$50; 60)] encompasses more objects (patients) than the attribute Age $\langle$50; 60) in the second rule. Our goal is to obtain as specific and detailed rules as possible and thus the first rule is omitted in the output because it is more general than the second rule.

In Figure 13, we can see a window with the output from one particular task solved by the Ac4ft-Miner. 8 action rules were found (the action rules are labelled as "hypotheses"). We can get a more detailed information about each rule in a separate window, including four-fold tables for the initial and final state and their representation in the form of diagrams.

Figure 13: Output in Ac4ft-Miner



5 Case study

In this chapter, I will present some of the examples based on the real data set ADAMEK. This set comprises the data from the cardiological patient; we have records regarding 1395 patients at our disposal. There are 37 columns which correspond to the properties of patients. On the basis of the columns, attributes are defined. The description of the data set is in Appendix 1.

5.1 Methodology

The goal is to find interesting patterns in the data matrix. I assume that an interesting pattern must include one or more risk factors of atherosclerosis.

5.1.1 Risk factors of atherosclerosis

According to [Euromise, c2010a], there are the following risk factors of atherosclerosis:

- Arterial hypertension – blood pressure over 140/90 mm Hg
- Hypercholesterolemy – cholesterol over 5 mmol/l
- HDL cholesterol – less than 1.1 mmol/l (men), less than 1.3 mmol/l (women)
- LDL cholesterol – over 3 mmol/l
- Hypertriglyceridemy – triglycerides over 2 mmol/l
- Smoking
- Overweight – BMI over 25
- Waistline – over 94 cm with men and over 80 cm with women
- Glycemy – over 6 mmol/l
- Diabetes mellitus
- Positive family case history

[Euromise, c2010b] adds following items:

- Sex – male
- Age over 45 for the male and 55 for the female
- The lack of physical activity

I assume that sex, age, positive family case history cannot be flexible attributes, other attributes can be considered either as a stable or as a flexible attribute.

5.1.2 Analytical questions

I will only study analytical questions in the following form:

Note: Text in green indicates an antecedent stable part, text in red indicates an antecedent variable part and text in blue indicates a succedent stable part.

Box 14: "Intuitive formulation"

> For which properties of patients (group 1) does the change of other properties (group 2) cause a change in the incidence of other properties (group 3)?

I call this form of formulation "intuitive formulation".

This formulation is then transformed into the form which I call “more formal formulation”. This is in the following form:

Box 15: “More formal formulation”

For which combinations of properties from group 1 does the relative frequency of combinations of properties from group 3 change by at least q by changing the combinations of properties from group 2.

In the formulation of analytical questions, the “combination of properties” is expressed in the following form: Property 1 and/or ... and/or Property n.

The expression “and/or” reflects the fact that the properties do not need to be present in the rule; it does not express conjunction or disjunction

Finally, there is the “formal definition” formulated as an input into the Ac4ft-Miner procedure:

Box 16: “Formal definition”

The formal definition consists of the definition of the antecedent stable part (group 1), the antecedent variable part (group 2), the succedent stable part (group 3) (the succedent variable part is not employed), and the definition of quantifier - the quantifier Founded implication $\Rightarrow^{I \lesssim F}_{q, B_I, B_F}$ which is defined as

$$\left| \frac{a_I}{a_I + b_I} - \frac{a_F}{a_F + b_F} \right| \geq q \wedge a_I \geq B_I \wedge a_F \geq B_F, \qquad \text{where } 0 < q \leq 1, B_I > 0 \text{ and } B_F > 0.$$

For details and examples of a definition of the input in the Ac4ft-Miner, see chapter 4.1.2.

5.1.3 Default settings

There are some default settings which are the same for each analytical question unless stated differently:

- Cedents are conjunctions of literals
- Minimal length of intervals and subsets is 1

5.2 Analytical questions

In this section, examples of using the Ac4ft Miner procedure are presented. In each example, at first the analytical question is stated. Then, the input parameters of a stable and flexible antecedent and a stable and flexible succedent in the form of table as well as quantifier are presented. The output consists of a number of hypotheses (i.e. rules) found, a number of verifications and the time needed to solve the task at PC with 2 GHz and 895 MB RAM. If applicable, some of the founded rules are presented together with their initial and final state. At the end, there is an interpretation of each rule and a comment. The text in *italics* specifies the formal input of a rule.

As mentioned above, the description of attributes and groups of attributes is to be found in Appendix 1.

There are five analytical questions:

Analytical question	Main idea
1	Too strict formal definition
2	A detailed demonstration of the rule founded and selected
3	Use of condition, all the rules presented in detail, various ways of presenting rules
4	Loose definition but no rules found
5	Loose definition, many insignificant rules found, long duration of the procedure

ANALYTICAL QUESTION 1

Intuitive formulation

For which personal measurements does a change of activities cause a change in the incidence of risk factors?

More formal formulation

For which combinations of BMI and/or Waist circumference and/or Hip circumference does the relative frequency of Hyperlipoproteinemy and/or Diabetes mellitus and/or Hypertension and/or Positive family case history change by at least 0.4 by changing Physical activity at work and/or Physical activity *(and the number of objects satisfying both initial antecedent and initial succedent is at least 40 and the number of objects satisfying both final antecedent and final succedent is at least 40)*.

INPUT

	Antecedent	Succedent
Stable part	BMI *(int. per 1, int. max length 6)* Waist *(int. per 5 cm, int. max length 3)* Hip *(int. per 5cm, int. max length 3)*	Hyperlipoproteinemy *(subset max length 1)* Diabetes mellitus *(subset max length 1)* Hypertension *(subset max length 1)* Positive family case history *(subset max length 1)*
Variable part	Physical activity at work *(subset max length 1)* Physical activity *(int. max length 2)*	

Quantifier	$\Rightarrow \underset{q, B_I, B_F}{I \lesssim F}$	$\left\lvert \frac{a_I}{a_I + b_I} - \frac{a_F}{a_F + b_F} \right\rvert \geq 0.4 \wedge a_I \geq 40 \wedge a_F \geq 40$

OUTPUT

Number of hypotheses (rules) found:	0	Number of verifications:	18,951,486
Duration at PC with 2 GHz and 895 MB RAM		0h 34m 45s	

Interpretation

There are no combinations of BMI and/or Waist circumference and/or Hip circumference for which the relative frequency of Hyperlipoproteinemy and/or Diabetes mellitus and/or Hypertension and/or Positive family case history change by at least 0.4 by changing Physical activity at work and/or Physical activity.

COMMENT

Both the absolute difference of the quantifier values, B_I, and also B_F were defined relatively high. This may be due to not finding any rules. It is possible to try to lower q, B_I and B_F and see if it brings any results (see Analytical question 2).

ANALYTICAL QUESTION 2

Intuitive formulation

For which personal measurements causes a change of activities a change in the incidence of risk factors?

More formal formulation

For which combinations of BMI and/or Waist circumference and/or Hip circumference does the relative frequency of Hyperlipoproteinemy and/or Diabetes mellitus and/or Hypertension and/or Positive family case history change by at least 0.3 by changing Physical activity at work and/or Physical activity *(and the number of objects satisfying both initial antecedent and initial succedent is at least 30 and the number of objects satisfying both final antecedent and final succedent is at least 30).*

INPUT

	Antecedent	Succedent
Stable part	BMI *(int. per 1, int. max length 6)* Waist *(int. per 5 cm, int. max length 3)* Hip *(int. per 5cm, int. max length 3)*	Hyperlipoproteinemy *(subset max length 1)* Diabetes mellitus *(subset max length 1)* Hypertension *(subset max length 1)* Positive family case history *(subset max length 1)*
Variable part	Physical activity at work *(subset max length 1)* Physical activity *(int. max length 2)*	

Quantifier	$\Rightarrow_{q, B_I, B_F}^{I \lesssim F}$	$\left\lvert \frac{a_I}{a_I + b_I} - \frac{a_F}{a_F + b_F} \right\rvert \geq 0.3 \wedge a_I \geq 30 \wedge a_F \geq 30$

OUTPUT

Number of hypotheses (rules) found:	80	Number of verifications:	46,648,080
Duration at PC with 2 GHz and 895 MB RAM:		1h 21m 54s	

One of the founded rules:

Waist (70; 85⟩ ∧ Hip ⟨95; 110) ∧ Physical activity at work (low → unemployed) ∧ Physical activity (low, medium → none, low) $\Rightarrow_{0.302,43,31}$ Hypertension (no)

Initial rule

Waist (70; 85⟩ ∧ Hip ⟨95; 110) ∧ Physical activity at work (low) ∧ Physical activity (low, medium) $\Rightarrow_{0.935,43}$ Hypertension (no)

	Hypertension (no)	¬ Hypertension (no)
Waist (70; 85⟩ ∧ Hip ⟨95; 110) ∧ Physical activity at work (low) ∧ Physical activity (low, medium)	43	3
¬ (Waist (70; 85⟩ ∧ Hip ⟨95; 110) ∧ Physical activity at work (low) ∧ Physical activity (low, medium))	908	441

Confidence = $\frac{a}{a+b} = \frac{43}{43+3} = 0.935$

Interpretation of initial rule

There are 43 patients with waist circumference between 70 and 85 centimetres, hip circumference between 95 and 110 centimetres, with low physical activity at work and low or medium physical activity, who represent 93.5 % of all patients with waist circumference between 70 and 85 centimetres, Hip circumference between 95 and 110 centimetres, with low physical activity at work and low or medium physical activity, who do not have hypertension.

Final rule

Waist ⟨70; 85⟩ ∧ Hip ⟨95; 110) ∧ Physical activity at work (unemployed) ∧ Physical activity (none, low) $\Rightarrow_{0.633,31}$ Hypertension (no)

	Hypertension (no)	¬ Hypertension (no)
Waist ⟨70; 85⟩ ∧ Hip ⟨95; 110) ∧ Physical activity at work (unemployed) ∧ Physical activity (none, low)	31	18
¬ (Waist ⟨70; 85⟩ ∧ Hip ⟨95; 110) ∧ Physical activity at work (unemployed) ∧ Physical activity (none, low))	920	426

Confidence = $\frac{a}{a+b} = \frac{31}{31+18} = 0.633$

Interpretation of final rule

There are 31 patients with waist circumference between 70 and 85 centimetres, hip circumference between 95 and 110 centimetres, unemployed, with none or low physical activity, who represent 63.3 % of all patients with waist circumference between 70 and 85 centimetres, hip circumference between 95 and 110 centimetres, unemployed, with none or low physical activity, who do not have hypertension.

Interpretation of the whole action rule

If the physical activity at work is changed from low to unemployed and physical activity from low or medium to none or medium among the patients with waist circumference between 70 and 85 centimetres and hip circumference between 95 and 110 centimetres, the incidence of patients not having hypertension decreases by 30.2 percentage points.

COMMENT

There were 80 rules found, one of which was presented in detail.

ANALYTICAL QUESTION 3

Intuitive formulation

For which personal measurements causes a change of activities a change in the incidence of risk factors in male patients?

More formal formulation

In case of male patients, for which combinations of BMI and/or Waist circumference and/or Hip circumference does the relative frequency of Hyperlipoproteinemy and/or Hypertension change by at least 0.3 by changing Physical activity at work and/or Physical activity *(and the number of objects satisfying both initial antecedent and initial succedent is at least 30 and the number of objects satisfying both final antecedent and final succedent is at least 30).*

INPUT

	Antecedent	Succedent
Stable part	BMI *(int. per 1, int. max length 6)* Waist *(int. per 5 cm, int. max length 3)* Hip *(int. per 5cm, int. max length 3)*	Hyperlipoproteinemy *(subset max length 1)* Hypertension *(subset max length 1)*
Variable part	Physical activity at work *(subset max length 1)* Physical activity *(int. max length 2)*	

Condition	Sex *(one category-(male))*	
Quantifier	$\Rightarrow^{I \lesssim F}_{q, B_I, B_F}$	$\left\lvert \frac{a_I}{a_I + b_I} - \frac{a_F}{a_F + b_F} \right\rvert \geq 0.3 \wedge a_I \geq 30 \wedge a_F \geq 30$

OUTPUT

Number of hypotheses (rules) found:	6	Number of verifications:	598,064
Duration at PC with 2 GHz and 895 MB RAM:		0h 1m 14s	

Note: In this analytical question, different approaches of presenting founded rules are introduced. All the rules founded are presented. The text in orange indicates the condition.

ACTION RULE 1

BMI (20; 26⟩ & Hip ⟨90; 105)
Physical activity at work (low) &Physical activity (low, medium) → Physical activity (none)
Hyperlipoproteinemy (no) 86.8% → Hyperlipoproteinemy (no) 54.4% / Sex (male)

Interpretation of the action rule

In case of male patients with BMI between 20 and 26 and with hip circumference between 90 cm and 105 cm, **IF** we change low physical activity at work and low or medium physical activity **TO** none physical activity, probability of not having hyperlipoproteinemy decreases **FROM** 86.8 % **TO** 54.4 %.

ACTION RULE 2

BMI (23; 28⟩ ∧ Hip ⟨75; 90) ∧ Physical activity (medium → none) $\Rightarrow_{0.311,30,30}$ Hyperlipoproteinemy (no) / Sex (male)

Interpretation of the action rule

In case of male patients with BMI between 23 and 28 and with hip circumference between 75 cm and 90 cm, **IF** we change physical activity **FROM** medium **TO** none, the incidence of not having hyperlipoproteinemy decreases **BY** 31.1 percentage points.

Initial rule

BMI (23; 28⟩ ∧ Hip ⟨75; 90) ∧ Physical activity (medium) $\Rightarrow_{0.811,30,30}$ Hyperlipoproteinemy (no) / Sex (male)

	Hyperlipoproteinemy (no)	**NOT** Hyperlipoproteinemy (no)
BMI (23; 28⟩ ∧ Hip ⟨75; 90) ∧ Physical activity (medium)	30	7
NOT **[** BMI (23; 28⟩ ∧ Hip ⟨75; 90) ∧ Physical activity (medium) **]**	305	241

Interpretation of initial rule

There are 30 male patients with BMI between 23 and 28, hip circumference between 75 and 90 centimetres with medium physical activity, who represent 81.1 % of all male patients with BMI between 23 and 28, hip circumference between 75 and 90 centimetres with medium physical activity, who do not have hypertension.

Final rule

BMI (23; 28⟩ ∧ Hip ⟨75; 90) ∧ Physical activity (none) $\Rightarrow_{0.500,\ 30}$ Hyperlipoproteinemy (no) / Sex (male)

	Hyperlipoproteinemy (no)	**NOT** Hyperlipoproteinemy (no)
BMI (23; 28⟩ ∧ Hip ⟨75; 90) ∧ Physical activity (none)	30	30
NOT **[** BMI (23; 28⟩ ∧ Hip ⟨75; 90) ∧ Physical activity (none)**]**	305	218

Interpretation of final rule

There are 30 male patients with BMI between 23 and 28, hip circumference between 75 and 90 centimetres with none physical activity, who represent 50 % of all male patients with BMI between 23 and 28, hip circumference between 75 and 90 centimetres with none physical activity, who do not have hypertension.

ACTION RULE 3
BMI (22; 28⟩ ∧ Hip (75; 90⟩ ∧ Physical activity (medium → none) $\Rightarrow_{0.300,32,32}$ Hyperlipoproteinemy (no) / Sex (male)
Interpretation of the action rule:
In case of male patients with BMI between 22 and 28 and with hip circumference between 75 cm and 90 cm, **IF** we change physical activity **FROM** medium **TO** none, the incidence of not having hyperlipoproteinemy decreases **BY** 30 percentage points.

ACTION RULE 4	
	IN PATIENTS WITH
Antecedent stable part	**BMI (20; 26⟩ & Hip ⟨90; 105)**
	WHO ARE
Condition	**Sex (male)**
	IF WE CHANGE
Ant. variable part - from	**Physical activity (none)**
Proposed change	TO
Ant .variable part - to	**Physical activity at work (low) & Physical activity (low, medium)**
	FOLLOWING PROPERTY
Succedent stable part	**Hyperlipoproteinemy (no)**
	WILL INCREASE IN ITS INCIDENCE BY
Difference of quantifier values	**0.324**
	MULTIPLE 100 PERCENTAGE POINTS

ACTION RULE 5
BMI (23; 28⟩ ∧ Hip (75; 90⟩ ∧ Physical activity (none → medium) $\Rightarrow_{0.311,30,30}$ Hyperlipoproteinemy (no) / Sex (male)
Interpretation of the action rule:
In case of male patients with BMI between 23 and 28 and with hip circumference between 75 cm and 90 cm, **IF** we change physical activity **FROM** none **TO** medium, the incidence of not having hyperlipoproteinemy increases **BY** 31.1 percentage points.

ACTION RULE 6
BMI (22; 28⟩ ∧ Hip (75; 90⟩ ∧ Physical activity (none → medium) $\Rightarrow_{0.300,32,32}$ Hyperlipoproteinemy (no) / Sex (male)
Interpretation of the action rule:
In case of male patients with BMI between 22 and 28 and with hip circumference between 75 cm and 90 cm, **IF** we change physical activity **FROM** none **TO** medium, the incidence of not having hyperlipoproteinemy increases **BY** 30 percentage points.

COMMENT
In this analytical question, the condition was used. Only the columns containing the attribute Sex (male) were included in the four-fold tables. The “opposite” rules were founded. Rules 1 and 4; 2 and 5; 3 and 6 are “opposite”. One rule from the pair suggests an action from the state X to the state Y while the succedent decreases in its incidence; the other rule from the pair suggests an action from the state Y to the state X while the succedent increases in its incidence. Both antecedent and succedent of the “opposite” rules are the same.

ANALYTICAL QUESTION 4

Intuitive formulation

For which age groups and/or which sex and/or which weight of patients does a change of BMI cause the cholesterol to change?

More formal formulation

For which combinations of Age and/or Sex and/or Weight does the relative frequency of Cholesterol change by at least 0.2 by changing BMI *(and the number of objects satisfying both initial antecedent and initial succedent is at least 10 and the number of objects satisfying both final antecedent and final succedent is at least 10)*.

INPUT

	Antecedent	Succedent
Stable part	Age *(int. per 5 years, int. max length 4)* Sex *(subset max length 1)* Weight *(int. per 5kg, int. max length 4)*	Total cholesterol *(int. per 0.5, int. max length 3)*
Variable part	BMI *(int. per 1, int. max length 6)*	
Quantifier	$\Rightarrow_{q,B_I,B_F}^{I \lesseqgtr F}$	$\left\lvert \frac{a_I}{a_I + b_I} - \frac{a_F}{a_F + b_F} \right\rvert \geq 0.2 \wedge a_I \geq 10 \wedge a_F \geq 10$

OUTPUT

Number of hypotheses (rules) found:	0	Number of verifications:	0
Duration at PC with 2 GHz and 895 MB RAM		7h 33m 3s	

Interpretation

There are no combinations of Age and/or Sex and/or Weight for which the relative frequency of Cholesterol changes at least by 0.2 as a consequence of changing BMI.

COMMENT

Although the q, B_I and B_F were defined relatively low, there were no rules found. The process of finding rules was relatively long due to many categories of attributes (BMI has 18 categories, Age 11 categories, Weight 12 categories, Cholesterol 9 categories) and also due to the maximum length of intervals of each attribute. There were, therefore, many possible combinations which have to be created and checked.

ANALYTICAL QUESTION 5

Intuitive formulation

For which risk factors does a change in cholesterol cause a change in the incidence of complaints?

More formal formulation

For which combinations of Hyperlipoproteinemy and/or Diabetes mellitus and/or Hypertension and/or Positive family case history does the relative frequency of Breathlessness and/or Claudication and/or Palpitation change by at least 0.2 by changing Total cholesterol and/or HDL cholesterol and/or LDL cholesterol and/or Triacylglycerols *(and the number of objects satisfying both initial antecedent and initial succedent is at least 10 and the number of objects satisfying both final antecedent and final succedent is at least 10).*

INPUT

	Antecedent	Succedent
Stable part	Hyperlipoproteinemy *(one category-(yes))* Diabetes mellitus *(one category-(yes))* Hypertension *(one category-(yes))* Positive family case history *(one category-(yes))*	Breathlessness *(one category-(due to the activity))* Breathlessness *(one category-(while inactive))* Claudication *(one category-(yes))* Palpitation *(one category-(yes))*
Variable part	Total cholesterol *(int. per 0.5, int. max length 2)* HDL cholesterol *(int. per 0.5, int. max length 2)* LDL cholesterol *(int. per 0.5, int. max length 2)* Triacylglycerols *(int. per 0.5, int. max length 2)*	
Quantifier	$\Rightarrow^{I \lesssim F}_{q, B_I, B_F}$	$\left\lvert \frac{a_I}{a_I + b_I} - \frac{a_F}{a_F + b_F} \right\rvert \geq 0.2 \wedge a_I \geq 10 \wedge a_F \geq 10$

OUTPUT

Number of hypotheses (rules) found:	437	Number of verifications:	742,569,534
Duration at PC with 2 GHz and 895 MB RAM:		22h 50m 31s Interrupted	

Interpretation:

There were 437 rules found. They are, however, not very significant – the highest confidence of initial rule is only 0.318.

COMMENT

In the input of the antecedent stable part and succedent stable part, in the definition of literals the option “one category” was used. This is because we want to find rules containing risk factors and complaints (we do not want to find rules containing, for example, Palpitation (no)). The generation and verification was interrupted after nearly 23 hours. This example shows that by defining the q, B_I and B_2 too low, we can obtain insignificant rules (they are true for very low number of patients), and it also demonstrates that the run of the Ac4ft-Miner is very long in this case.

6 Methodology of use of the Ac4ft-Miner for doctors

This chapter summarises the experience with the employment of the Ac4ft-Miner and proposes the methodology of use. Note that no precise methodology could be created due to the large number of options that Ac4ft-Miner offers. Thus, the following paragraphs present mere recommendations for the use of the tool which could help in the further research.

1. Follow the CRISP-DM Methodology – set the aim of the KDD project and the data mining tasks precisely because, as my experience has proved, this is crucial for the formulating of analytical questions
2. Group the attributes – it is useful to arrange the attributes into groups and work with them in the form of a group while formulating analytical questions (examples of the groups of attributes can be found in Appendix 1)
3. Pay attention to the formulating of analytical questions – it has proved wrong to formulate analytical questions too specific because it usually leads to no results (as shown in analytical question 4), and, therefore, the use of the groups of attributes is more appropriate. By formulating the analytical questions on a more general level, the probability of getting unexpected and interesting results is higher.
4. However, consider the balance between the generality and the specificity – it is not good to formulate the analytical question in *too* general terms. This often leads to finding more rules, these are however insignificant and the generation and verification take a long time (as shown in analytical question 5).

The methodology also depends on the nature of the data used as the input.

Conclusions

The Ac4ft-Miner is a data mining tool that enables to mine for rules suggesting an action. It is a part in the whole KDD process which I presented with the help of CRISP-DM methodology. Its first phase is the business understanding, where in particular the principal aim of the KDD project from the managerial point of view is specified. The second phase, the data understanding, serves as the first insight into the available data. In the third phase, the data preparation, the data relevant to the data mining task is selected and adjusted to the requirements of the employed data mining tools. In the fourth phase, the modelling, the data mining tools are used, and in the fifth phase, the evaluation, the results are compared to the business goals set in the first phase. If they are not sufficient, it is imperative to return to the previous phases. This possibility represents one of the main features of the CRISP-DM methodology and can be applied in all its phases. In the last phase, the deployment, the results are applied to the practical management.

For the representation of the data, the Ac4ft-miner uses Boolean attributes. They are created from the database columns. For each Boolean attribute, it is possible to determine whether it is true or not for each particular object (row) in the database. One of the concepts incorporated into the Ac4ft-Miner is association rules. The "classical" association rules introduced by [Agrawall et al., 1993] are in the form: antecedent implies consequent. The enhanced association rules which are generated by the GUHA method are more general. They enable to define more types of relationships. These rules are then used to form G-action rules. G-action rules were inspired by action rules introduced by [Ras, Wieczorkowska, 2000], but again they are more general. Furthermore, they suggest an action that brings an advantage. They define flexible attributes, which can be modified by the change of behaviour, and stable attributes, which cannot be changed. G-action rule can be expressed by two association rules – one for the initial state (before the change) and one for the final state (after the change).

There are various options of defining a set of relevant rules in the Ac4ft-Miner. The output consists of all the rules which are true in a data matrix and which are prime. In the case study, I apply the Ac4ft-Miner on the real medical data set ADAMEK, which consists of records regarding cardiological patients. The goal of my case study was to find interesting patterns related to atherosclerosis. I assume that an interesting pattern must contain one of the risk factors of atherosclerosis. I define a three-level formulation of analytical questions: "intuitive formulation", which is most likely understandable to non-experts in the data mining; "more formal formulation", which contains some of the aspects related to the data mining task; and "formal definition", which equals the input in the Ac4ft-Miner. I restrict my research only to the use of the quantifier *Founded implication* $\Rightarrow^{I \leq F}_{q, B_I, B_F}$.This implies that I try to find such action rules whose initial rule has its confidence higher or lower than the final rule by at least q and number of objects satisfying both initial antecedent and initial succedent is at least B_I and number of objects satisfying both final antecedent and final succedent is at least B_F.

Using the examples of analytical questions, I demonstrate the high variability of the defining input in the Ac4ft-Miner and basic principles of the procedure. The task solving lasts from a few minutes to several hours depending on the parameters of the input. Some of the rules founded are presented in detail to show how the action rules have been created; the rules are interpreted in a way which is easy to understand. The methodology of use for doctors is in the form of advices for the use of the tool which should help in the further research that is needed. This is because of the high complexity of the procedure, which does not allow formulating general conclusions usable in the methodology.

The principal aims of this thesis were to:

1. provide an overview of the Ac4ft-Miner procedure
2. present examples of the use of the Ac4ft-Miner procedure on the real medical data set ADAMEK
3. convey the experience obtained
4. create a methodology of use of the Ac4ft-Miner for doctors

The aim 1 has been achieved as demonstrated in chapter 3, where the theoretical background of the procedure is presented; moreover, in chapter 4, where the input and output of the procedure is introduced. The aim 2 has been achieved as obvious from chapter 5, where examples of analytical questions and their visualisation are presented. The aim 3 has been achieved as shown in chapter 6. The aim 4 has been achieved partly as may be seen in chapter 6 – this is due to the high complexity of the procedure.

The output of this thesis is a coherent text with lot of examples separated from the continuous text; so the reader familiar with a particular topic can skip the examples and proceed to the next issue. Further result of this thesis is an outline of the graphical presentation of analytical questions. Both the examples and the graphical presentation will be used further in the SEWEBAR project of which this thesis is one part.

The thesis can be further elaborated in particular by giving precision to the methodology of use. Moreover, it is imperative to accomplish more tasks while using different types of data employed as the input. This would provide an appropriate prerequisite for creating more sophisticated methodology.

Bibliography

Agrawall, R.; Imielinski, T.; Swami, A. 1993. Mining Association Rules between Sets of Items in Large Databases. IN: Proceeding of the 1993 ACM SIGMOD Conference Washington DC, USA, May 1993. [electronic document]. Available from http://rakesh.agrawal-family.com/papers/sigmod93assoc.pdf

Anand, S.; Buchner, A.; Bell, D.; Hughes, J. 1996. Towards real-world data mining. [electronic document] Available from http://mcbuchner.com/HTML/Research/PDF/PAKM96b.pdf

Berka, P. 2003. Dobývání znalostí z databází. 1. vydání Praha : Academia 2003

Chapman, P.; Clinton, J.; Kerber, R.; Khabaza, T.; Reinartz, T.; Shearer, C.; Wirth, R. 2000. CRISP-DM 1.0. Step-by-step data mining guide. [electronic document]. Available from http://www.crisp-dm.org/CRISPWP-0800.pdf

Euromise c2010a. Co je to ateroskleróza? [online].[cited 2010 June 8]. Available from http://www.euromise.cz/health/preventive_cardio/ateroskleroza.html

Euromise c2010b Kardiovaskularni onemocneni [online].[cited 2010 June 8]. Available from http://new.euromise.org/czech/kardio/kardio/frames2.htm

Fayyad, U.; Piatetsky-Shapiro,G.; Padhraic, S. 1996. From Data Mining to Knowledge Discovery in Databases [electronic dokument]. 1996. Available from http://www.kdnuggets.com/gpspubs/aimag-kdd-overview-1996-Fayyad.pdf

Kodym, J. 2007. Třídy SD4ft pravidel [master thesis]. Univerzita Karlova v Praze. Matematicko-fyzikální fakulta. 2007

Ras, Z. 2007. Action rules IN: 3rd Polish and International PD FCCS'2007 [electronic document]. 2007

Ras, Z.; Wieczorkowska, A. 2000. Action-Rules: How to Increase Profit of a Company IN: D.A. Zighed, J. Komorowski, and J. Zytkow (Eds.): PKDD 2000, LNAI 1910, pp. 587-592 Berlin : Springer Verlag, 2000.

Rauch, J. 2005. Systém LISp-Miner. Stručný popis určený pro posluchače kurzu Metod zpracování informací [electronic document]. Version 20 November 2005.

Rauch, J. 2009. Zadání úloh na téma „Dobývání znalostí z databází“ pro předmět 4IZ 210 Zpracování informací a znalostí vyučovaný na KIZI FIS VŠE Praha v letním semestru 2008/2009 [electronic document]. Praha, April 2009

Rauch, J.; Šimůnek, M. 2005. GUHA Method and Granular Computing. In: HU, X et al. (eds.). Proceedings of IEEE conference Granular Computing 2005.

Rauch, J.; Šimůnek, M. 2009. Action Rules and the GUHA Method: Preliminary Considerations and Results. Praha 14.09.2009 – 17.09.2009. In: Foundations of Intelligent Systems. Berlin : Springer Verlag, 2009, pp. 76–87. ISBN 978-3-642-04124-2. ISSN 1867-8211.

Rauch, J.; Šimůnek, M. c2010. Action Rules and the GUHA Method - Preliminary Considerations and Results [PowerPoint presentation].[cited 2010 March 6]

Rauch, J.; Tomečková, M. et al. 2009. ADAMEK – popis dat (verse III) [electronic document]. EuroMISE centrum. Praha, April 2009.

SEWEBAR (SEmantic WEB and Analytical Reports). c2010 [online]. [cited 2010 June 19]. Available from http:// sewebar.vse.cz

STULONG: Details about project STULONG. c2010 [online]. [cited 2010 June 8]. Available from http://euromise.vse.cz/stulong-en/metodika/index.php?page=metodika

Šťastný, D. 2009. Knowledge Processing within the GUHA Method [bachelor thesis]. Vysoká škola ekonomická v Praze. Fakulta informatiky a statistiky. 2009

XLMiner: Association Rules – Introduction, c2010 [online]. [cited 2010 March 12]. Available from http://www.resample.com/xlminer/help/Assocrules/associationrules_intro.htm

Appendix 1 – Description of the data set ADAMEK

This data description is adopted from [Rauch et al., 2009]. The data is of a Czech origin. Data was collected in two cardiological ambulancies of Euromise in Prague and Čáslav. The data set represents records regarding 1395 patients. From 37 columns of the data set, attributes have been created (see the table below). They were arranged into groups to make the formulation of analytical questions easier. It is important to note that other groups for a particular data mining task can be created as well.

There are three types of attributes:

- **N – nominal** – categories are presented in the table, the order of categories does not matter
- **O – ordinal** - categories are presented in the table, the order is important
- **R – rational** – from the data in the database, intervals of various length can be created and these are used in a data matrix

Group			Attribute				
No.	**Abb.**	**Name**	**No.**	**Name**	**Name of column in database**	**Type**	**categories**
1	OSB (3)	Personal data	1	Age	Vek	R	
			2	Sex	Pohl	N	woman / man
			3	Ambulance	M	N	Čáslav / Praha
2	SCA (5)	Social anamnesis	4	Marital status	Rod	N	divorced / single / widow(er)/ married
			5	Lives alone	Sam	N	yes / no
			6	Education	Vzd	O	primary / secondary / university
			7	Mental strain	PsychZat	O	none / low / medium / high
			8	Smoking	Koureni	N	Non-smoker/ exsmoker/ occassional smoker / smoker
3	AKT (2)	Activities	9	Workload	FyzZat	N	unemployed / none/ low / medium / high
			10	Physical activity	TelAkt	O	none / low / medium / high
4	MRY (7)	Personal measurements	11	Weight	Hmotnost	R	
			12	Height	Vyska	R	
			13	Waist circumference	Pas	R	
			14	Hip circumference	Boky	R	
			15	BMI	BMI	R	
			16	WHR	WHR	R	
			2	Sex	Pohl	N	woman / man
5	RFK (4)	Risk factors	17	Hyperlipoproteinemy	Hlp	N	yes / no
			18	Diabetes mellitus	DM	N	yes / no
			19	Hypertension	HT	N	yes / no
			20	Positive family case history	Ra_f	N	yes / no
6	OBT (5)	Complaints	21	Breathlessness	Dusnost	O	none / due to the activity/ while inactive
			22	Chest pain	Bolesthrud	O	none / due to the activity/ while inactive

Group			Attribute				
No.	**Abb.**	**Name**	**No.**	**Name**	**Name of column in database**	**Type**	**categories**
			23	Palpitation	Palpitace	N	yes / no
			24	Swelling	Otoky	N	yes / no
			25	Claudication	Klaudikace	N	yes / no
7	KTL (2)	Blood pressure	26	Systolic	ST	R	
			28	Diastolic	DT	R	
8	EKG (3)	EKG	29	Pulse rate	TepFrek	R	
			30	PQ_int	PQ_int	R	
			31	QRS_int	QRS_int	R	
9	CHL (4)	Laboratory Cholesterol	32	Total Cholesterol	Chol	R	
			33	HDL Cholesterol	HDL	R	
			34	LDL Cholesterol	LDL	R	
			35	Triacyglycerols	Tgl	R	
10	GKM (2)	Laboratory GKM	36	Glycemy	Glyk	R	
			37	Urine acid	Kmoc	R	