

**CZECH TECHNICAL UNIVERSITY IN PRAGUE
FACULTY OF ELECTRICAL ENGINEERING**

DEPARTMENT OF CYBERNETICS

DIPLOMA THESIS

**Intelligent analysis of data from obstetrics module
of hospital information system**

Prague 2012

Ondřej SUCHÝ

DIPLOMA THESIS ASSIGNMENT

Student: Bc. Ondřej S u c h ý

Study programme: Cybernetics and Robotics

Specialisation: Robotics

Title of Diploma Thesis: Intelligent Analysis of Data from Obstetrics Module of Hospital Information System

Guidelines:

The aim of the thesis is to find interesting features obtained from the hospital information system module of the obstetric ward in Teaching Hospital Brno.

1. Familiarize yourself with the issue of data mining in a medical environment and available tools for the computerized data processing.
2. Familiarize yourself with data provided from the hospital information system (HIS) Teaching Hospital in Brno. Describe the data.
3. Use association rules to find interesting signs implying selected important states during or immediately following birth.
4. Use at least one other method to check found features.
5. Document the results of experiments carefully and, if possible, interpret them.
6. Describe the most technical problems which were encountered during work, suggest their solution in the future.

Bibliography/Sources: Will be provided by the supervisor.

Diploma Thesis Supervisor: Ing. Václav Chudáček, Ph.D.

Valid until: the end of the winter semester of academic year 2012/2013


prof. Ing. Vladimír Mařík, DrSc.
Head of Department




prof. Ing. Pavel Ripka, CSc.
Dean

ZADÁNÍ DIPLOMOVÉ PRÁCE

Student: Bc. Ondřej S u c h ý

Studijní program: Kybernetika a robotika (magisterský)

Obor: Robotika

Název tématu: Inteligentní analýza dat z porodnického modulu nemocničního informačního systému

Pokyny pro vypracování:


Cílem práce je nalézt zajímavé příznaky získané z porodnického modulu nemocničního informačního systému FN Brno.

1. Seznamte se s problematikou dolování dat v medicínském prostředí a s dostupnými nástroji pro její počítačové provedení.
2. Seznamte se s daty, poskytnutými z nemocničního informačního systému (NIS) FN Brno. Data popište.
3. Pomocí asociačních pravidel najděte zajímavé příznaky implikující vybrané důležité stavy v průběhu či bezprostředně následně po porodu.
4. Takto nalezené příznaky ověřte ještě minimálně jednou metodou.
5. Výsledky svých experimentů důkladně dokumentujte a pokud je to možné, interpretněte.
6. Popište největší technické problémy na něž jste v průběhu práce narazil, navrhněte jejich řešení do budoucna.

Seznam odborné literatury: Dodá vedoucí práce.

Vedoucí diplomové práce: Ing. Václav Chudáček, Ph.D.

Platnost zadání: do konce zimního semestru 2012/2013


prof. Ing. Vladimír Mařík, DrSc.
vedoucí katedry




prof. Ing. Pavel Ripka, CSc.
děkan

Abstract

Association rules and their application are of the main concern of this diploma thesis. The aim is to obtain useful information from real obstetrics database of singleton deliveries. Data mining softwares (RapidMiner, Lisp-Miner and Orange) are used to help with this problem. State of the art data mining of obstetrics field is stated as well.

Our obstetrics data consist of approximately 500 potential features. Most of them are redundant, therefore it was possible to reduce the set to 106 attributes based on literature and with the aid of obstetricians of FN Brno hospital even more to 54 significant features.

This thesis investigates outcome of newborn, influences of parameters on delivery by Caesarean section and macrosomia of newborn via association rules with Fisher quantifier. Among the most significant results are: support medicine such as oxytocin does not influence positive outcome of newborn, hypoxia of fetus mostly leads to Caesarean section, the more previous deliveries, the better result in spontaneous delivery and that body mass index has impact on macrosomia.

Resulting features are also evaluated by classification accuracy, sensitivity and specificity of Random Forest classification.

Among the biggest problems encountered were data preparation, low computing power and interpretation of association rules – common problems in such a research task.

Keywords

Data mining, Data analysis, Association rules, Artificial intelligence, CTG, Hypoxia

Number of pages, tables, figures and appendices

Number of pages:	84
Number of tables:	6
Number of figures:	23
Number of appendices:	4

Abstrakt

Tato diplomová práce se zaměřuje na asociační pravidla a jejich použití. Cílem je získat užitečné informace ze skutečné databáze porodnických záznamů, které se týkají doby před porodem nebo po porodu (porody jednoho novorozence). Používám nástroje pro dolování dat (RapidMiner, Lisp-Miner a Orange), abych tento problém zvládl. Zmiňuji také aktuální články o dolování dat v porodnictví.

Naše porodnická data se skládají přibližně z 500 možných příznaků. Většina z nich je nadbytečná, proto je bylo možné zredukovat na 106 příznaků na základě literatury. S pomocí porodníků z Fakultní nemocnice Brna se datová sada dále zredukovala dokonce na 54 důležitých příznaků.

Tato diplomová práce zkoumá výsledek novorozence, císařský řez a makrosomii plodu pomocí asociačních pravidel s Fisherovým kvantifikátorem. Mezi nejvýznamnější výsledky patří: podpurná medicína jako je např. oxytocin neovlivní pozitivní výsledek novorozence, hypoxie plodu ve většině případů vede k císařskému řezu, více předchozích těhotenství přispívá k současnému spontánnímu porodu a index tělesné hmotnosti ovlivňuje makrosomii.

Příznaky z asociačních pravidel jsou dále vyhodnoceny pomocí přesnosti klasifikace, senzitivity a specificity klasifikační metodou „Random Forest“.

Největší zjištěné problémy jsou příprava dat, malá výpočetní síla počítačů a interpretace asociačních pravidel – běžné problémy v takové výzkumné úloze.

Klíčová slova

Dolování dat, Analýza dat, Asociační pravidla, Umělá inteligence, KTG, Hypoxie

Počet stran, tabulek, obrázků a příloh

Počet stran:	84
Počet tabulek:	6
Počet obrázků:	23
Počet příloh:	4

Declaration

I hereby declare that I have completed my Diploma thesis independently and that I have used only cited sources (literature, projects, SW etc.) listed in part “Bibliography“.

In Prague on:

.....

Ondřej Suchý

Prohlášení

Prohlašuji, že jsem svou diplomovou práci vypracoval samostatně a použil jsem pouze podklady (literaturu, projekty, SW atd.) uvedené v příloženém seznamu.

V Praze dne:

.....

Ondřej Suchý

Acknowledgement

I would like to thank to Ing. Václav Chudáček, Ph.D. (CTU in Prague, K13133) for supervision and any help related to completion of my diploma thesis.

Special thanks belong to my family, which supports me a lot within my studies.

Content

Content.....	8
List of used symbols and abbreviations.....	11
List of Tables.....	12
List of Figures.....	13
INTRODUCTION.....	16
1 Introduction and Aims of the Diploma thesis.....	17
THEORETICAL PART.....	19
2 Data mining.....	20
2.1 Terminology.....	20
2.2 Historical background.....	22
2.3 Statistics, Data mining, Knowledge Discovery in Databases.....	22
2.4 Data pre-processing (Data preparation).....	24
2.5 Data mining techniques.....	24
2.5.1 Nearest neighbour.....	25
2.5.2 Cluster analysis (Clustering).....	25
2.5.3 Decision trees.....	26
2.5.4 Artificial neural networks.....	27
2.5.5 Association rules.....	28
2.6 Visualization methods.....	31
2.7 Results validation.....	33

3 Software tools for data analysis.....	34
3.1 SchemaSpy.....	34
3.2 RapidMiner.....	35
3.3 Lisp-Miner.....	36
3.4 Orange.....	37
3.5 Weka.....	38
4 Overview of related research in medical domain.....	40
4.1 Associations, Attributes.....	40
4.2 Similar datasets, Methods.....	41
PRACTICAL PART.....	43
5 CTU Database.....	44
5.1 Concept.....	44
5.2 Details.....	45
5.3 Example of reading data from the database.....	45
5.4 Data description and explanation.....	46
6 Experiments.....	50
6.1 Data preparation.....	50
6.1.1 Manual selection of attributes.....	50
6.1.2 Physiological limitations.....	51
6.1.3 New attributes, Categorization, Adjusting names of classes and attributes.....	52
6.1.4 Datasets creation and completion.....	54
6.2 Results.....	56
6.2.1 Association rules, Question 1.....	57
6.2.2 Association rules, Question 2.....	59
6.2.3 Association rules, Question 3.....	59
6.2.4 Association rules, Question 4.....	60
6.3 Discussion.....	61
6.3.1 Question 1.....	61

6.3.2 Question 2.....	65
6.3.3 Question 3.....	66
6.3.4 Question 4.....	67
7 Evaluation of results via other data mining method.....	68
8 Encountered problems and possible future solution.....	71
CONCLUSION, BIBLIOGRAPHY, APPENDICES.....	73
9 Conclusion.....	74
Bibliography.....	77
List of appendices.....	80
Appendix 1.....	81
Appendix 2.....	82
Appendix 3.....	83
Appendix 4.....	84

List of used symbols and abbreviations

HIS	...	FN Brno Hospital information system
CTU	...	Czech Technical University in Prague
DM	...	Data mining
CTU Database	...	Database of Biodat Research Group of CTU
BMI	...	Body mass index
pH	...	pH measure (potential of hydrogen)

List of Tables

Table 1	4-fold table with letters that stand for a frequency of instances in an antecedent or a succedent. (Table taken from source [3]).....	29
Table 2	Annotations of obstetrics data. (Classes and annotations are in Czech language; “NE“ means “no“, “ANO“ is “yes“, “vetsi...” means “higher than“ and “mensi...” is “lower than“)......	47
Table 3	Selected attributes of obstetrics data. (Classes and attributes are in Czech language; “ne“ means “no“, “ano“ is “yes“, “vetsi...” means “higher than“ and “mensi...” is “lower than“)......	47
Table 4	Limitations of attributes that could contain errors.....	52
Table 5	Classification of 5-nearest neighbour classifier and Random Forest classifier with 10 individually developed trees. Lines Marked in grey are classifications of entire dataset and white lines represent classifications of selected attributes.....	70
Table 6	Classifications of Random Forest classifier with 10 individually developed trees. Selected attributes and annotations represent classifications of relevant attributes of association rules from Chapter 6.....	70

List of Figures

Figure 1	RadViz example with classes “Proteas” (blue), “Resp” (red) and “Ribo” (green). (Taken from website of source [19]).....	32
Figure 2	Example of naive Bayesian nomogram of heart disease data. (Taken and adjusted from source [19]).....	33
Figure 3	Example of SchemaSpy command to connect to CTU Database and to create the schema for the web browser.....	35
Figure 4	Example of association rules results produced by Lisp-Miner with Fisher quantifier. Lenght of antecedent is 2. I do not provide explanation here, because most of hypothesis are not valuable. Besides, the summation of results and interpretation is given in next chapters.....	37
Figure 5	This SQL script selects attributes "ic_chor", "mens_od", "teplota" from tables "osobni_udaje2", "anamneza_lekar" and "anamneza_sestra" respectively. It uses alias names "ou2", "anl", "ans" to shorten the length in large SQL script. To connect attributes from different tables by corresponding number of medical record "ic_chor", commands "JOIN" and "USING" are necessary.....	46
Figure 6	RapidMiner process, which creates entire datasets into Weka files from CTU Database.....	54
Figure 7	RapidMiner process, which creates datasets of weight of newborn into Weka files and Microsoft Access Database file from entire dataset. Operator “Read ARFF” is renamed to “arff ic_chor” and operator “Select Attributes” to “sel-att”. 55	
Figure 8	Results of Lisp-Miner task with Fisher quantifier, antecedent of length from 1 to 1 and succedent is class “vetsi7” of apgar sum in 5th minute.....	58

Figure 9	Results of Lisp-Miner task with Fisher quantifier, antecedent is selected attributes from Figure 8 of length from 1 to 6 and succedent is class “vetsi7“ of apgar sum in 5th minute.....	58
Figure 10	Results of Lisp-Miner task with Fisher quantifier, antecedent is selected attributes (not presented in diploma thesis) of length from 1 to 6 and succedent is class “mensi7_15“ of pH of newborn.....	59
Figure 11	Results of Lisp-Miner task with Fisher quantifier, antecedent is selected attributes (not presented in diploma thesis) of length from 1 to 5 and succedent is class “SC“ of method of childbirth.....	60
Figure 12	Results of Lisp-Miner task with Fisher quantifier, antecedent is attribute “Porody_celk“ and the succedent all classes of method of childbirth.....	60
Figure 13	Results of Lisp-Miner task with Fisher quantifier, antecedent is selected attributes (not presented in diploma thesis) of length from 1 to 10 and succedent is class “vetsi4000“ of weight of fetus.....	61
Figure 14	Simple association rule of streptococcus with 4-fold table and some of all possible Lisp-Miner metrics.....	63
Figure 15	4-fold table of association rule of antecedent streptococcus and oxytocin. Confidence is 99.5% and coverage 3.1%.....	63
Figure 16	4-fold table of association rule of antecedent streptococcus, oxytocin and pregnancy week. Confidence is 99.7% and coverage 3.0%.....	63
Figure 17	4-fold table of association rule of antecedent streptococcus, oxytocin, pregnancy week and number of previous childbirths. Confidence is 100% and coverage 1.1%.....	64
Figure 18	Classification process of Orange software. The process opens dataset with proper annotation, selects attributes according to association rules and do the classification of Random Forest and Nearest neighbour classifiers.....	69

Figure 19	Orange software - default setup of a) 5-Nearest neighbour classifier and b) 10 individually developed trees in Random Forest classifier.....	69
Figure 20	Example of a simple pseudo decision tree. Circles are nodes and predictions are in squares. Classes are above branches of the decision tree.....	81
Figure 21	Example of relations among tables of CTU Database generated by SchemaSpy software. Understandably, the table personal data “osobni_udaje” has the most relations, because it contains specific attributes used in the rest of tables.....	82
Figure 22	Example of possible data preparation and data transformation in RapidMiner 5.2. The process read from CTU database according to settings in “Parameters“ window. Further settings of the database and other operators is not shown. Then it selects attributes for annotation “ANOTACE_ctg-_spatny“ and set it as a label (special attribute) from regular one. It filters missing values of the annotation and write Weka “.arff“ file in nominal and in numerical form (“Nominal to Numerical“ operator is used).....	83
Figure 23	Example of evaluation of the result from Chapter 6 in Orange software. Classification target is annotation “ANOTACE_apgar_suma_5“ with class “vetsi7“. It shows classification accuracy, sensitivity and specificity of Nearest neighbour (kNN) and Random Forest classifiers. 10-fold cross validation is used.....	84

INTRODUCTION

1 Introduction and Aims of the Diploma thesis

Data analysis is an inspection of data, which provides previously unknown significant information or knowledge.

It is a big issue these days in many fields, e.g. business, medical or social domain etc. Mainly due to the fact, that the inspection of data can provide us useful information in a reasonable time frame.

Data analysis is built on gathering whichever historical data related to the research problem and data mining algorithms that do the inspection of data. Sometimes it can be useful to collect data that does not relate to the research problem. They can help us to refill the information underlying the research problem.

There is huge amount of algorithms, techniques and software tools that assist us to get the best results (prediction of future). It is necessary to apply proper procedure for any dataset to obtain fine prediction of future states, relations and the like. This is exactly the challenge for anybody who is involved in data mining process.

The aim of my diploma thesis is similar. I should have inspected real obstetrics data from FN Brno Hospital information system (HIS) and to find valuable attributes.

Obstetricians in FN Brno hospital may use these attributes as a support in their decision within pregnancy, childbirth, delivery or states following the delivery.

I divide my diploma thesis into 4 parts (introduction, theoretical part, practical part and to conclusion, bibliography and appendices part). Chapters 2, 3, 4 are in theoretical part and Chapters 5 to 8 are in practical part of this thesis.

Chapter 2 is a general introduction to data mining and data mining techniques.

Chapter 3 provides information about freeware data mining softwares.

Chapter 4 is the state of the art overview of attributes, that influence childbirth, and of data mining databases in obstetrics field or medical domain.

I describe HIS database and its data in **Chapter 5**.

Chapter 6 is the actual data preparation and association rules extraction part. I use association rules to find relations (patterns) among attributes that could influence fetus, newborn or gravida. Then I interpret resulting valuable association rules.

Chapter 7 is the evaluation of results (attributes) obtained in Chapter 6.

Chapter 8 deals with problems encountered within this diploma thesis (e.g. software, hardware problems).

THEORETICAL PART

2 Data mining

This chapter is a general introduction to data mining.

I explain terms used in data mining (DM), historical background, data mining itself, differences between statistics, DM and knowledge discovery in databases. Then I shortly describe data pre-processing.

I focus on the basics of common DM techniques and visualization methods, e.g. nearest neighbour, decision trees or association rules.

I also mention validation of results.

2.1 Terminology

This part serves as an explanation of basic terms related to DM and knowledge discovery in databases. I use most of terms in my diploma thesis.

Attribute (also known as “Feature“) is a certain quality inherent in data, e.g. attendance to school, hair colour, electricity consumption and so on. In a database, attribute is a column of the table.

Class is a property or value of one attribute, for example if our attribute is electricity consumption, the attribute may contain two classes – the first one is “high“ (means high electricity consumption), the second “low“ (means low electricity consumption). In such a case, the class is in a nominal form. Another class may comprise of numerical values (i.e. “1“, “2“, “3“, “4“ etc.), which describe the attribute. Then the class is in a numerical form. Nominal or numerical forms are also discrete forms. The example of the numerical class of the attribute “hair color“ is “1“ (stands for blonde hair), “2“ (stands for brown hair), “3“ (stands for ginger hair). It is obvious that this kind of description can cause problems upon explaining the results to natural speech. Further important forms of classes are boolean form (i.e. “t“ or “f“, stands for true or false) and binary form (i.e. “0“ or “1“), real form (i.e. decimals) or continuous form (i.e. description by a curve).

Instance is one record of the attribute. In a database, it is a row or cell under the heading of the attribute. One of the classes of the attribute is stored here.

Data are certain facts. Data may be records of measurement, results of calculations etc.
[1] Database is an example of well-ordered data with defined access.

Dataset is a chosen subset of attributes and instances from a database. Data pre-processing is worked out over the dataset before an actual DM. Dataset can also be seen as already prepared set of attributes after pre-processing for easy use of DM algorithms.

Feature extraction is the selection of attribute or attributes that have the biggest impact on a surveyed target (attribute or class).

Pattern is a characteristic behaviour in a dataset. A combination of attributes (classes) is typical for a target attribute (target class). Mostly, such characteristic behaviour is hidden. People are very curious about it, what if it contains important and valuable information. Therefore data miners search for patterns in datasets very often.

Prediction is an expectation of a future state of the attribute based on recorded data. The typical task is to find a certain value of the attribute through knowledge of other attributes. Prediction is divided into classification and regression. **Classification** is the task, where the result is a discrete value. Discrete value is one of the classes of the attribute.[1] For example combination of attributes “intensity of barking“ and “height“ provides the result “hound“ (the type of dog). **Regression** is the task, where the result is a real number.[1] For instance the temperature is measured all the time - within rain, cloudy or sunny weather and season. For example, in case of combination of “rain“, “sunny weather“ and “summer“, the regression provides expectation of resulting temperature “26.5“ degrees Celsius.

Indication is a retrieval of unusual patterns presented in a dataset, e.g. indication of fault state of a technological system. Then we can prepare for the situation when fault state occurs.[1]

Description is a retrieval of new knowledge and new regularities that contributes to the development of a specific branch.[1]

Model is a representation of relations among predictors (input attributes) and target attribute or class. It provides the lowest error in a prediction. The process of finding and creating the model is called modeling.[12]

Data mining software is every software intended to perform data mining over certain dataset. DM software implements DM techniques and turns them into practical application. It can be used as a tool that do the analysis, categorization or relation retrieval. It can assist

people to make the best decision according to prediction and relevant dataset.

2.2 Historical background

From the beginning of computers, there was a necessity to store data, mainly for reporting in business or plots in engineering. With the development of information society, there is a massive collection of data and classical statistical techniques became insufficient or less comfortable while answering more in quantity and difficulty questions.[12]

Retrospective questions such as “How much products was sold in last 2 years?” transformed into prospective questions such as “What is going to happen when the price change? Why is it so?”.[12]

Answering these questions preceded long research in three domains – artificial intelligence, machine learning and statistics, from which new algorithms for DM were invented. Afterwards, step by step improvements in technology helped to push ahead new DM algorithms. Technology improvements are e.g. data collection changed to data warehousing, data access transformed to on-line data delivery at multiple levels and noneffective computers evolved into powerful multicore and multiprocessor computers or computer grids.[12]

In recent years, based on such technology, DM algorithms proved their maturity in a sense of reliability, understandability and higher performance than classical statistics.[12] Probably, the crucial moment for a huge deployment of DM these days is the performance and computer strength.

2.3 Statistics, Data mining, Knowledge Discovery in Databases

A difference between statistics and DM or DM and knowledge discovery in databases is very tiny. Certain DM techniques grew up from statistics. Regarding DM and knowledge discovery in databases, they are even considered to be equal by most people, which is not fully true as explained later.

Statistics is purely based on mathematics. It is a significant branch of mathematics, because it arose from real world business or biological problems. Therefore statistics is useful,

if we want to make a decision, where the risk or uncertainty are present.[12]

Statistics and DM are based on the same ideas, e.g. independence, causality, defined target, clean data or good validation. They may be used for similar purposes as prediction or classification discovery. Main distinction is that statistics is less resistant to faults in data and it has to be understood before an application. The more and cleaner data is then a decision is better, but with statistician expert supervision and the decision stays on a responsible person. The second thing is that statistics focus on collection, counting and summarization of data. Therefore it provides high level view on a data, which is the most valuable for reporting.[12] Another fundamental applications of statistics are hypotheses testing and a detection of dependence among attributes.

DM is a part of computer science, which bridges applied statistics with artificial intelligence (provides mathematical background) and it also works as intersection of these fields with machine learning (technical background) and database systems.[13]

The term DM comes from the parallelism with era of gold diggers, who mined for a gold, very important ore for a lot of people. In DM, the hidden gold is identified with the hidden interesting information.[12]

More formally, DM is a process, where discovery of new typical behaviour previously hidden in data is fundamental. Usual task of DM is automated or semi-automated analysis of large datasets.[13] DM helps to analyze data from different angles, categorizing and summarizing them to significant and useful information. Then the typical behaviour may be used for more accurate prediction.

In a broader context, DM is a step in a **knowledge discovery in databases** process, because solely DM is inconsequential without additional steps. A simplified process knowledge discovery in databases is:

- Data pre-processing,
- DM,
- Validation of results.[13]

Nevertheless, there exists a lot of modifications of the process according to the requirements of data miners or companies. At the turn of 20th and 21st centuries, there were efforts to define standards of the process, but without broader success. Nevertheless, the process called CRISP-DM is well known. It starts with a problem and data

understanding, continues with data preparation, DM (modeling, evaluation) and finishes by deployment of results. Another usual variation of the knowledge discovery in databases process can be feature (attribute) selection, data pre-processing and transformation, DM, results evaluation and interpretation.[13]

Why is it actually good to know about the process? According to me, the only answer is that it can help people, who start with DM and who follow these steps to spare time and to obtain exploitable results easily.

In my diploma thesis, I also apply additional steps of knowledge discovery in databases. Data pre-processing (see Chapter 6.1 - Data preparation), results interpretation (Chapter 6.3) and results validation (Chapter 7) are part of this thesis.

2.4 Data pre-processing (Data preparation)

Data pre-processing is a preparation of data before DM.[13] The reason for this step is that data usually does contain errors, mistakes, bad format for DM or apparently irrelevant attributes. With the help of data pre-processing, following analyses are simpler and faster. Data assey is a detailed descripton of dataset and its attributes and classes. It belongs to data pre-processing as well because it can identify what data contains, what problems are in data and howto unify data.[1] Usual data pre-processing methods are data cleansing and discretization of data. Data cleansing is a removing instances with noise or missing data. [13] Discretization of data is a categorization to classes, e.g. people 0 to 20 years old are category one, with age 21 to 40 are placed under category two etc.

2.5 Data mining techniques

Data mining technique is a general method howto perform DM task. Every task needs a specific aproach and own data preparation.

There exists several DM techniques, which we consider as basic techniques and from which there are many adjusted or derived techniques.[12] That basic techniques I describe in this part, e.g. nearest neighbour, cluster analysis, decision trees, artificial neural networks and association rules. Classification and regression tasks are ommited because I have already

explained their principle in the previous part under term “Prediction”.

I pay more attention to association rules because they are the main concern of my diploma thesis.

2.5.1 Nearest neighbour

Nearest neighbour technique is one of the oldest DM prediction technique. To obtain the finest prediction, the algorithm search for the most similar records of input attributes in the historical dataset and provides the closest value or class of the target attribute for this combination of input attributes.[12]

Despite the fact that the principle is easy, this technique is not possible to use in everyday life because of inaccurate data or of a very specific situations, which occur rarely. The prediction would mislead us, if we rely only on one nearest example.[12]

Therefore “k-nearest neighbour” technique evolved, which relies on the number “k” closest examples, e.g. “k” equals 10, the algorithm search for 10 closest examples presented in a dataset and provides prediction in a form of an average value or of the major class according to occurrence of 10 nearest target classes. It was proven that simple nearest neighbour technique has lower prediction accuracy than k-nearest neighbour technique.[12]

The nearness is determined by the measurement of distance between or among matching records. The measurement of distance is also very useful while articulating a confidence of the prediction.[12]

2.5.2 Cluster analysis (Clustering)

The reason for cluster analysis is to get a “big picture“ of data with similar properties. [12] Next utilization is the anomaly (outlier) detection in data.[14]

The principle of cluster analysis is to search through a dataset and find similar objects (instances of attributes), which are then grouped. One group is called a cluster. Every cluster has more similar objects then the others.[14]

Source [12] provides nice example of clustering – laundry separation. People usually separate their laundry before the washing. Let say that our dataset is laundry, the clusters are

white, coloured, dry cleaning, soft wash etc. This selection is based on the “behaviour” of laundry (white sustain 80 degrees celsius, coloured less) while washing in the washing mashine because we do not want to damage it. Some pieces of laundry may be for example white with coloured stripes. It is difficult to decide whether to group them into white or coloured laundry. Probably, better decision would be to group them into coloured laundry because dying of all white laundry is worse than dying white parts of some pieces of striped laundry. Of course, it depends on other qualities such are popularity of laundry, money or another qualities.

Cluster analysis is an explorative tool of a dataset, but different DM algorithms separate dataset to different clusters. Here, DM researcher plays fundamental role because his or her goal influences the result and afterwards the interpretation.[14]

The similarity of objects is detected by distance (hierarchical clustering), statistical distribution (distribution-based clustering), density (density-based clustering) or central vectors (centroid-based clustering, objects are grouped to the nearest cluster centre).[14]

2.5.3 Decision trees

Decision tree is a set of rules that represent a dataset.[2] The representation in a visualization form is a hierarchical structure that resembles to a tree. Nodes are attributes and values or classes separate the node to branches. One value or one class is next to the branch, which connects two nodes of different levels of hierarchical structure. At lowest level of every branch is a leaf, it means the prediction of the class of surveyed attribute.

The example of a simple decision tree is in Appendix 1 with visualization in Figure 20.

Decision tree may completely describe the dataset. In ideal case, the sequence of decisions gives us correct recommendation for any prediction of the class in the dataset. In real life, more complex and descriptive decision tree model does not mean that it performs better in the prediction. The representation of data may be approximated, too. Therefore the prediction of certain values or classes is not exact. In any case, an explanation of the decision tree is intuitive and easy. However, certain rules must not make obvious sense because there is something underlying it.[2]

To build the decision tree, there exists two base concepts – general-to-specific search and specific-to-general search. General-to-specific search starts with the general problem

and it tries to split the problem into the particles. The later works the other way round, it start with the simple particles of data and it tries to make a general description of the problem.[2]

To obtain superior decision tree, test of quality is necessary. There are test of particular attribute, test of attribute value to a constant, test attribute to a function or a comparison of multiple attributes. Upon the creation of a decision tree, pruning of the tree is executed to avoid overfitting of the model on test data.[2] Due to pruning, decision trees are smaller and less complex.

Decision trees are originally meant as complete DM process, from hypothesis generation to validation of model. Thus they can handle datasets without pre-processing and they are highly successful in many applications of prediction and exploration problems. They can even do pre-processing or selection of attributes for other DM techniques (e.g. neural networks). Decision trees are not good in applications, where an equation describe the target attribute.[12]

2.5.4 Artificial neural networks

Artificial neural networks arose from the attempt to mimic human brain in structure and function, i.e. detecting patterns, making prediction and learning. It is a DM algorithm to create predictive models from large datasets. The idea is that a set of simple units (artificial neurons) can solve difficult problems, which are even more complicated than that simple units.[12]

Artificial neural networks consist of nodes and links. Input nodes are input attributes, output node is a prediction attribute. In case of multiple output nodes, output nodes are the best attributes. This structure is used as automated feature extraction. Attributes have normalized classes in range 0 to 1. Between input and output layer, there is usually one hidden layer of nodes that perform actual learning of artificial neural networks. However, the structure may consist of many hidden layers of nodes. Links are connections among nodes of different layers. Every node of one lower layer of nodes is connected to every node of one upper layer of nodes.[12]

The principle of learning is the same as working in an army. In the army there is one general that has the final decision (same as a target prediction attribute). The general has advisors in a lower layer and these advisors have own advisors in the next lower layer etc.

If the general makes a good decision, the advisors that gave a good advice gain higher confidence from their superiors to give them future advice. The advisors with a bad advice lose some amount of the confidence of their superior to give them future advice. To transform the army topic into artificial neural networks, the advisors are nodes (attributes or combination of attributes) and the confidence of a future advice is done by weighting of any link in the structure. Higher weight of a link means higher confidence and the other way round.[12]

The artificial neural networks adjust weights of links one record at a time. Firstly the model is built from a training set. Weights of links of the model remain the same on new data until an error is made. Afterwards weights are modified to provide right prediction for the next record.[12]

The advantage of artificial neural networks is the high accuracy of a prediction and the application on diverse problems. The disadvantage is the explanation of the result (numerical values) and of the hidden layers, data pre-processing and skilled DM personality to design and work with a network.[12]

Note that it is sometimes better to use support vector machine technique, which needs less attributes with the same or greater accuracy.[2]

2.5.5 Association rules

Association rules are rules that can associate any attribute class in a dataset and they can describe or predict different things. This is the main distinction from classification rules that predict only one class for one specified attribute. In a dataset there are usually a few association rules 100% correct and many more less correct. Most of 100% correct are based on a small number of instances. In the rest of association rules, anomalies or missing values do occur and therefore lower correctness is present.[2]

The standard example to understand association rules is the task of analysis of shopping basket. A seller wants to sell as much goods as he or she can. The store has only a list of sold goods from individual sales. The history proved that certain sort of goods is sold in relation with another sort. For example simple association rule may be:

- **IF sold coffee AND milk THEN sugar** is sold as well.

Probably, people buy this combination for a coffee break. One of possible interpretation of such a relation for a seller is that he or she should reorder counter to let people mention certain goods, e.g. sugar next to milk. Another interpretation could be that the store should concentrate on a certain segment of goods and omit insignificant goods, e.g. lemonade.

The left side of association rule (coffee AND milk) is called an antecedent, the right side is called a succedent (sugar). A notation of this association rule is also “Antecedent => Succedent”. [3]

The connection between the antecedent and succedent is a logical implication (THEN, =>) or equivalence (<=>). In this case it is the logical implication. The connection among attributes of antecedent or succedent is logical conjunction (AND, ^) or disjunction (OR, v). In some cases, a negation (\neg) of attribute is used, too.

In a large dataset, there can be a huge amount of association rules and it is very problematic to target best ones. Thus limitations by metrics are used, e.g. minimum accuracy or instances etc. [2]

Metrics are based on a 4-fold table (see Table 1), which shows a frequency of instances from a dataset in a relevant antecedent and succedent combinations.

Table 1: 4-fold table with letters that stand for a frequency of instances in an antecedent or a succedent. (Table taken from source [3])

	Succedent	Not Succedent	Sum of row
Antecedent	a	b	$r = a + b$
Not Antecedent	c	d	$s = c + d$
Sum of column	$k = a + c$	$l = b + d$	$n = a + b + c + d$

Chosen metrics are defined according to source [3]. Explanations are according to different sources. Formulas and explanation are given here:

- $Support = \frac{a}{n}$: It is a percentage of instances in entire dataset that fulfils the presented association rule. [15]

- $Confidence = \frac{a}{r}$: It is an accuracy of the rule.[12] It states that an association rule is correct in $\frac{a}{r}$ percents of cases. Another possible explanation is that it is a probability of finding succedent under condition of antecedent.[15]
- $Coverage = \frac{a}{k}$: It is a percentage of instances in a succedent that fulfils the presented association rule. I put here one note, it is obvious from equations that there is a slight difference between support and coverage. However, source [12] consider the coverage and support equal. Thus it is good to understand these two terms as defined rather than expressing them as equal.
- $Leverage = \frac{1}{(n*a)-(r*k)}$: It compares instances additionally covered in case of entire association rule over instances covered in case of independence of both sides of an association rule. Leverage equal to number 0 states the independence of an antecedent and succedent. Higher number of leverage states the dependence of an antecedent and succedent.[3] Maximum value of leverage should be 0.25.[4]
- $Lift = \frac{a*n}{r*k}$: It determines amount, about which it improves the precision of a correct prediction of a succedent.[3] Lift equal number 1 states the independence of an antecedent and succedent. The dependence of an antecedent and succedent is given by values far from number 1. Higher dependence expresses more interesting rule.[4]
- $Conviction = \frac{r*l}{b*n}$: It is the ratio of an incorrect prediction of an independent attributes and sides of an association rule divided by an observed incorrect prediction of the association rule.[15] Conviction equal number 1 states the independence of an antecedent and succedent. The dependence of an antecedent and succedent is given by values far from number 1.[4]

To search for association rules, confidence and support are valuable metrics. High confidence gives us highly accurate rules that imply the strong relationship and should be exploited everytime. High support gives us rules, which occur most of the time in a dataset.[12] If possible, the first step is to apply support on a dataset and the second step is to apply defined confidence range on pre-selected dataset by the support.[15]

Despite the metrics, manual selection of association rules is necessary to interpret meaningful rules.[2]

2.6 Visualization methods

Visualization of data is any graphical form of data.[5] Visualization methods are ideas or implementations how to graphically represent data of DM task.

Recently, a concern of visualization transformed from simple informative (only data presentation) to interpretational state.[5] Thus visualization methods are also used in explorative analysis to get to know the data. Via different methods we are able to find anomalies (errors), unusual situations,[2] attribute interactions and obtain global or quick view on a specific DM task.[5]

I omit basic visualization methods such as scatterplot, statistical distribution or decision trees. Besides, good explanation and outline of visualization methods are in Doctoral Thesis of Lenka Nováková (source [5]).

I mention RadViz and Nomogram visualization methods, which can represent multiple dimensions (multiple attributes) in one two-dimensional plane.

RadViz works in three steps. It normalizes data records to range 0 to 1 of all attributes related to the problem. It places these attributes symmetrically on a unit circle (here attributes are called dimensional anchors) and draws instances of attributes into the circle according to calculation of position.[5]

RadViz example is presented in Figure 1.

The idea underlying the positioning of instances goes from forces in physics. Forces affect each other. If we bind them into one system, after while they originate one equilibrium state (“not moving point”). In RadViz, forces are dimensional anchors, equilibrium state is determined by values of attributes and an instance is placed just into the equilibrium.[19]

Highest values of instances are close to dimensional anchors, e.g. high values of attribute “spo5 11” are typical for class “Ribo” and other attributes have significantly lower values for this class. The next explanation of presented Figure 1 is that high values of attributes “heat 20” and “diau f” are typical for class “Resp”. [19]

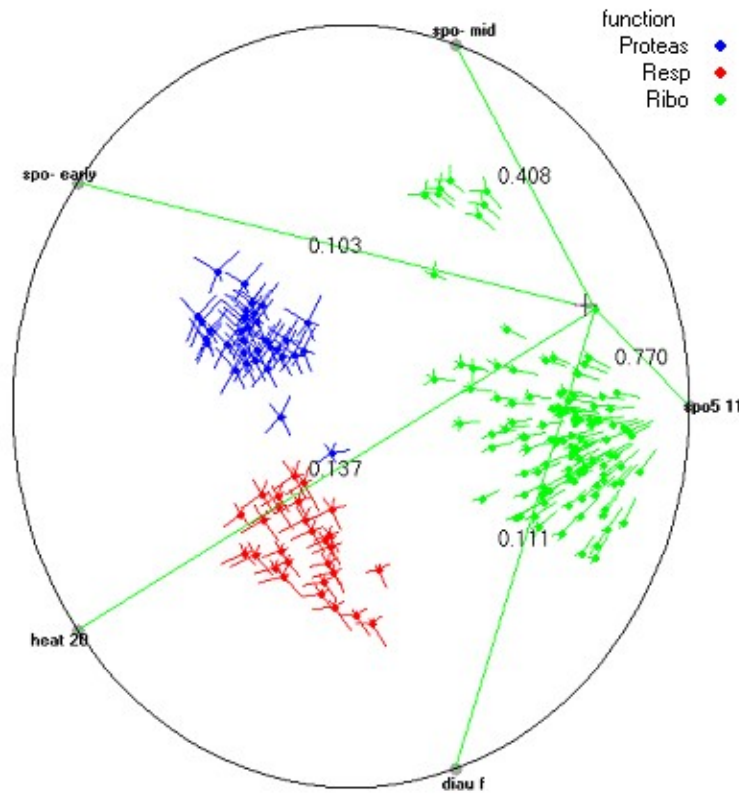


Figure 1: RadViz example with classes “Proteas” (blue), “Resp” (red) and “Ribo” (green). (Taken from website of source [19])

Nomogram is a representation of attributes based on calculation of a metric log odd ratio (logarithmic scale of an association) of each value of each attribute. It is possible to plot nomogram of different tasks. I explain nomogram for prediction of heart disease data as presented in [19] and adjusted for Figure 2.[19]

In this example we want to predict having heart disease and inspect what causes it. Having heart disease is our target class. Nomogram of naive Bayesian classifier shows attributes “gender“, “chest pain“, “rest SBP“, “cholesterol“, “max HR“. Vertical dashed line represent zero influence of an attribute on target class. The left side from dashed line is negative influence on target class and the right side is positive influence. Blue point is actual log odd ratio value of an attribute (in Figure 2 stated as “Points“ at top), e.g. class “male“ of an attribute “gender“ has +0.4 points or value of log odd ratio. This class contributes to having heart disease. Attribute „rest SBP“ has -0.51 points. Thus it contributes to not having heart disease.[19]

Sum of all points of attributes is -0.58 and it is listed at bottom of Figure 2. This value corresponds to 32% probability (indicated as 0.32 “Log OR Sum“ in Figure 2)

that the combination of attributes gives us right prediction of having heart disease.[19]

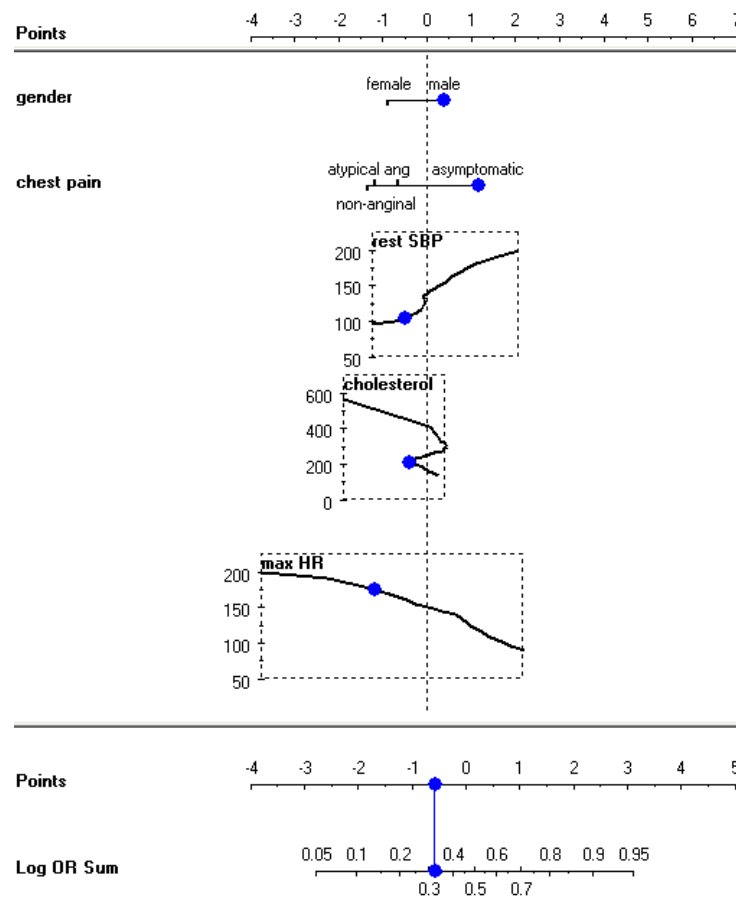


Figure 2: Example of naive Bayesian nomogram of heart disease data. (Taken and adjusted from source [19])

2.7 Results validation

Results validation is a verification of produced patterns by DM algorithms. It is necessary step because not all patterns are valid all the time. The problem is with overfitting, it means that if DM algorithms learn itself by training on entire dataset, some of produced patterns are not valid on the new dataset. The solution is to split dataset to training set for learning and to test set for an evaluation of patterns before they are applied to new dataset. The evaluation of a pattern is a comparison of prediction produced by the pattern and of the class present in test set. When the evaluation is highly accurate (correct most of time), the pattern is interpreted. Otherwise there is space for improvement in a pre-processing or in selection of DM algorithm.[13]

3 Software tools for data analysis

There exists huge amount of software tools for data mining. I concentrate on freeware, tools for multiple operating systems (mainly Windows and Linux) and tools that are easy or relatively easy to learn in a reasonable time.

Therefore I choose SchemaSpy, RapidMiner, Lisp-Miner, Orange and Weka. The description of these tools are given in Chapters 3.1 – 3.5.

3.1 SchemaSpy

Schema Spy is a very handy tool to create web pages of a database schema. The schema is then used as listing of entire database or to see relations among database tables in a web browser. It provides all relations, type and size of attributes.

SchemaSpy helps me to familiarize with the database of CTU Biodat Research Group (CTU Database).

This tool supports many database management systems such are PostgreSQL, MySQL, Oracle, DB2 etc. It is written in Java under LGPL 2.1 license and it analyses metadata of the database schema. To visualize the database in the web browser, former installation of Graphviz software tool is necessary. In downloaded “schemaSpy_5.0.0.jar” file, there is also lack of JDBC driver and “.properties” file of the corresponding database. JDBC driver and “.properties” file help to access the database and to perform transactions over the database.

Natively, SchemaSpy is controlled via command line, see Figure 3. SchemaSpyGUI is better option for people, who begin with SchemaSpy.

The command of Figure 3 opens file “schemaspy.jar”, finds PostgreSQL JDBC4 driver of database type “pgsql”. Then the database name, host, port and user is specified to connect to the proper database. Further, additional properties of PostgreSQL are specified. The password (fictional in this case), CTU Database schema “public” and output folder “iga_schema” is in the end of the command.

The software is possible to download from source [16].

```

java -jar "C:\schemasp\schemaspy.jar" -dp
"C:\schemasp\drivers\postgresql-9.1-901.jdbc4.jar"
-t pgsqql -db brnodata_step5 -host
biodata.felk.cvut.cz -port 5432 -u brnodata_reader
-connprops "C:\schemasp\properties\pgsqql.properties"
-p "password" -s public -o "C:\schemasp\iga_schema"

```

Figure 3: Example of SchemaSpy command to connect to CTU Database and to create the schema for the web browser.

3.2 RapidMiner

RapidMiner 5.2 is probably the best open-source data mining tool these days. It can directly access/export data from/to different storages (databases, .csv, .xml, .arff files etc.) and it can handle large datasets. There are options of data transformation, visualization, evaluation, modeling and even for text processing or web mining.

RapidMiner runs on main operating systems and there is a possibility to run it in java mode, too. Via java mode, computer memory may be allocated according to our needs.

Primarily, I use this tool for data preparation and data transformation.

To start with RapidMiner, I would recommend “RapidMiner Video Tutorials” on “Rapid - I” (source [17]) or “YouTube” websites.

The graphical user interface of this software consists of windows, from which windows “Operators”, “Process”, “Parameters” are the most important. Useful are also windows “Repositories”, “Problems” or “Help”. I do not use the rest of windows.

The concept of RapidMiner is based on a visual programming and process creation. The work with this software is intuitive while creating “light” processes. To evaluate data or create a model, additional knowledge of this tool is a big advantage.

“Operators” represent functions and they form a process under “Process” window. I can start the process and wait for the result in “Result Overview” window, if operators are

connected properly. If not, “Quick Fixes“ option under “Problems“ window can help me to handle the problem.

“Parameters“ are used to set properties of operators. “Help“ is the description of operators. “Repositories“ is the list of folders and RapidMiner files.

Example of a process and RapidMiner environment is shown in Appendix 3.

3.3 Lisp-Miner

Lisp-Miner is based on GUHA method (General Unary Hypotheses Automaton). I use this software to obtain association rules. It can be used for classification, clustering or outlier detection as well.

Lisp-Miner website (source [18]) describes and explains the method and the work with Lisp-Miner software. Lisp-Miner is possible to download from Lisp-Miner website in packed format as “.zip“ archive.

Software is designed for operating system Windows and it runs without installation. Lisp-Miner consists of small applications and every application opens in a special window. Applications may be initialised separately or from the main application “LMAdmin.exe“.

Data are prepared for the application “4ftTask.exe“ in application “LMDataSource.exe“. “4ftTask.exe“ produces hypotheses (association rules) according to our task setup. Application “4ftResult.exe“ makes for viewing results.

Data are processed only in a form of Microsoft Access database (.mdb files). Two database files are necessary. The first one is our own database with prepared data. It has to contain one attribute, which is marked in application “LMDataSource.exe“ as the database primary key and which distinguishes individual instances. In my case, the database primary key is attribute "ic_chor". The second database is called “Lisp-Miner metabase“ and it serves to save setup of tasks and results. Lisp-Miner metabase for experiments is a clone of the prepared file “LMEEmpty.mdb“. According to a Lisp-Miner naming convention, both databases should have a similar name and the metabase should contain the prefix “LM“.

Attributes are selected from a database before experiments in “LMDataSource.exe“. These attributes are then used in application “4ftTask.exe“, where I set one class of one attribute as succedent and the rest of attributes as antecedents. Parameters such are base, class,

length of association rule, handling missing values or quantifiers are optional. Base is the minimum value that antecedent implies succedent. Quantifiers are used as metrics for evaluation. They are based on 4-fold frequency table (e.g. Support, Founded Implication) or statistic methods (Fisher, Chi-square). The settings “Condition“ is possible to use when the attribute is surely associated with the succedent. The experiment is run after setup by button “Generate“.

The results are in a form of association rules. Antecedents are logically connected by pre-defined conjunctions. Association rules are easily readable and understandable, however it depends on the naming of attributes and theirs classes. Figure 4 shows example of results.

Actual group of hypotheses: All hypothesis

Hypotheses in group: 1841 Shown hypotheses: 1841 Highlighted: 0

Nr.	Id	Conf	Hypothesis
1	867	0.994	Doba1trv_epoch_int(0h) & Poloha_plodu_kod_vys(koncem_panevrim) *** ANOTACE_zpusob_porodu(SC)
2	860	0.993	Doba1trv_epoch_int(0h) & Krev_ztrata_int(500_1000) *** ANOTACE_zpusob_porodu(SC)
3	1132	0.992	Dr_vedl_doba_2_corr_hypoxie(ano) & Hmot_nov_int(1500_2500) *** ANOTACE_zpusob_porodu(SC)
4	1181	0.991	Dr_vedl_doba_2_corr_primamisc(ano) & Plod_voda_kod_vys(cista) *** ANOTACE_zpusob_porodu(SC)
5	1184	0.991	Dr_vedl_doba_2_corr_primamisc(ano) & Tt_dokonceny36_int(vetsi36) *** ANOTACE_zpusob_porodu(SC)
6	1180	0.991	Dr_vedl_doba_2_corr_primamisc(ano) & Krev_ztrata_int(1_500) *** ANOTACE_zpusob_porodu(SC)

Figure 4: Example of association rules results produced by Lisp-Miner with Fisher quantifier. Length of antecedent is 2. I do not provide explanation here, because most of hypothesis are not valuable. Besides, the summation of results and interpretation is given in next chapters.

3.4 Orange

Orange is very easy and still powerful software for data mining and machine learning. I use this tool for data visualization and classification. It can also be used for data preparation, feature extraction, regression, unsupervised learning or association rules.

Orange runs on main operating systems – Windows, Linux, Mac OS. The necessity to start the software is Python. Python is installed within installation of Orange under Windows.

This software is based on visual programming. Every function has one icon called “widget“. Widgets are placed in the main menu into groups according to the functions, e.g. widgets “File“, “Save“, “Data Table“ are placed in the group “Data“. Series of connected widgets forms a process. The process runs automatically (or manually; based on setup) after

any connection between widgets or after any change of settings of the widget. It might be useful for small datasets to run it automatically, but I would mark it as a drawback while working with large datasets, especially in a classification process. Additionally, there is a possibility to create scripts in Python to extend the function of Orange.

I would point out functions as “Rank“, “Outliers“, “Radviz“ and “Nomogram“. Rank selects attributes according to ranking method, for example Information gain, Gini index etc. Outliers function detects outliers in data based on distance metric between examples. Radviz selects attributes that separate the annotation and it visualizes results. Nomogram represents classification model and it can sort attributes according to importance to a prediction class of the annotation.

Orange is specific by its releases. The latest version is the development version 2.5, which is updated by a stable version every day. Newer version may have new function or new help of function.

There are no tutorials that would guide us through a programme. Examples of a process are shown in Orange website (source [19]). An explanation of functions are also in Orange website or as help in the software.

3.5 Weka

Weka is very simple, though it has implementations of many algorithms. Weka has preprocessing, classification, clustering and visualization capabilities, too.

Weka is developed in Java and runs on all main operating systems.

Applications “Simple CLI“, “KnowledgeFlow“, “Experimenter“ and “Explorer“ forms Weka. Simple CLI is a command line. KnowledgeFlow is a visual programming environment. Experimenter is an extension of Explorer because it can run several algorithms or several datasets in one setup. It shows results with comparison, which algorithm is statistically better than another. Explorer is an environment to work with one dataset.

The environment of Explorer is divided into tabs that perform all functions from pre-processing to visualization as scatter plot. Further self-explanatory settings are under every tab. “Choose“ button is used to set algorithm or DM technique. Double-click on selected algorithm or method invokes properties that are possible to change.

Regarding tutorials, there exists “Weka Manual“ for different releases of Weka that are possible to download from official website (source [20]).

4 Overview of related research in medical domain

I use online available databases and search engines, that CTU has access to, to do a research in a medical domain, especially in obstetrics field. I would like to gain information about significant attributes that influence pregnancy or delivery and associations between them. Further, I search for similar databases or datasets as Biodat Research Group of CTU provided to me, i.e. similar to HIS and related to pregnancy domain. In the later case, I expect aid in a form of how people solve the problem, which DM methods they use or how they make data pre-processing over the real dataset.

I utilize electronic articles from SpringerLink search engine. Search keywords in my case are “large dataset pregnancy”, “large dataset pregnancy, data mining”, “association rules pregnancy”.

4.1 Associations, Attributes

Processes within pregnancy and childbirth are not fully described and understand.[6] Source [6] focus on stillbirth and preterm birth, because new knowledge may lead to proper and effective intervention. It also describes spontaneous term parturition because it can help to uncover pathological mechanism of preterm delivery.

Normal parturition is characterised by physical activation of uterine components, but a preterm parturition results from pathological states in any step of pregnancy. Any perturbation in implantation of embryo leads to abortion, stillbirth and preeclampsia. Preeclampsia is a significant cause of a preterm birth.[6]

Early preterm parturition (24 to 32 week) is influenced by infection or inflammation, late preterm birth (32 to 37 week) is associated with stress and uterine overdistension. The article states that 62% women experience rapid preterm delivery, if amniotic fluid is infected in preterm period. Sterile amniotic fluid leads to rapid preterm birth in 13% of cases. Bacterial vaginosis increases risk of preterm birth, amniotic infection and low weight of fetus. Another factor for poor fetus outcome and preterm birth has an environment and pollution.[6]

Source [7] targets its work on body mass index (BMI) of gravida, weight gain

of gravida and birth weight of fetus. In research part, it states that birth weight is essential for infant mortality, childhood development and adulthood health, e.g. low birth weight can cause type-2-diabetes or heart disease. High birth weight can cause delivery complications (dystocia of shoulders).

According to this study, smoking, diabetes or race of gravida is not associated with birth weight. In opposite, pre-pregnancy BMI is strongly associated with birth weight, but the reason of such relation is hidden for this study. Another attributes that influence birth weight are preeclampsia and nulliparity. Weight gain is considered as mediator between input value of BMI and outcome of birth weight of fetus. However some investigators deny the dependence. The study deals with this difference that there is dependence between BMI and weight gain only in case of macrosomnic babies (over 4000 grams).[7]

BMI was also studied in the article of source [8], but they had inspected data of nulliparous women, which deliver only one newborn. Results are that higher BMI (e.g. obese women, BMI in range 30 to 34.9) increases risk of having following problems – preeclampsia, hypertension, complicated pregnancies and more interventions. What's more, increasing BMI is associated with induction of birth or Caesarean section. Low BMI should eliminate certain complications.[8]

4.2 Similar datasets, Methods

I did find only several large datasets that resemble to HIS and relates to obstetrics field. Most of them use statistical evaluation of attributes and logistic regression method to determine the association among attributes. Underneath, I mention datasets of sources [8, 9, 10, 11].

Dataset in source [8] includes 24 241 nulliparous women delivering singleton babies. Authors consider this dataset as one of the largest of this topic. The dataset was extracted from Aberdeen Maternity and Neonatal Databank of 200 thousands pregnancy records. Missing data were excluded or equally distributed to BMI categories. BMI is calculated from weight and height of gravida. Authors use statistical tests for attribute extraction with probability of 95% to be statistically significant, e.g. chi-square test for categorical attributes or ANOVA (analysis of variance) for continuous attributes. Risk of birth complications evaluates odd ratio metric.[8]

Missouri maternal dataset (United States of America) comprises of approximately 232 thousands of singleton births. To induce a model of diabetes, authors concentrate on methods as in previous case. Chi-square statistical test is used to choose attributes and logistic regression is used to evaluate risk for diabetes according to BMI.[9]

Danish National Birth Cohort dataset has approximately 92 thousands records of singleton birth. Feature extraction is based on BestFirst search algorithm. Attributes are added one at a time to form different sets of independent attributes. Sets of attributes are then evaluated. The one, which provides best prediction for the outcome is kept. Associations of attributes and risk for preterm delivery evaluate regression analysis.[10]

The last dataset that I want to mention is United States Nationwide Inpatient Sample (vide source [11]). This dataset consists of almost 8 million hospital records, but not related to pregnancies. It is a valuable source, because it covers field of solving DM problems of imbalanced data. Imbalanced data are in sense that one or more categories of an attribute have considerably higher amount of instances than other category or other categories.

Authors find random sub-sampling method (use equal number of instances from each class) with random forest DM method very effective to classify highly imbalanced data. Random forest is aggregated learner (aggregate results of multiple classifiers). It randomly searches for input attributes to determine the split of a decision tree. The split is determined by Gini index (measure of impurity, lower impurity means better split). Random forest can handle missing data and it can determine the importance of attributes, too.[11]

The performance of random forest is measured by ROC curve (plot of sensitivity) and AUC (average area under the curve). In the study, random forest outperforms support vector machine, boosting (set of weak learners produce strong learner) and bagging (model averaging) methods.[11]

PRACTICAL PART

5 CTU Database

Data are stored in many places in our computers. It would be impossible to make a proper data analysis over a non-organized data.

In many cases, data are organized in tables in files such as Microsoft Excel and the like. These files are possible to use for data mining, but they have certain limitations.

Therefore a database is the best option to organize data. It has advantages such as access by many persons at the same time, one location of all information, easy search through data or security. Data are also stored in tables, but every table has its primary key and thus it is possible to connect different attributes from different tables.

Here, I focus on description of obstetrics database and data, which are obtained from HIS, organized in CTU Database and maintained by Ing. Miroslav Burša, Ph.D.

Modern data mining techniques are based on visual techniques, approximation and usage of large datasets. To spare time and ease the job of people, who perform data analyses, computers and special softwares are used. Suitable software tools are listed in Chapter 3. that help me to carry out analysis over CTU Database.

5.1 Concept

CTU Database is created in PostgreSQL object-relational database management system. This concept of database management system tries to combine the advantages of both management systems (relational and object-oriented).

Relational concept organises data in tables and it uses standards known as SQL3 to define and transact data. Object-oriented concept allows to do programming via object - oriented programming languages (C++, Java etc.) over database to create richer data structures known as abstract data types.[21]

To be more specific, a table in the relational concept stores data, which are somehow related, e.g. personal data. Every table contains rows and columns. Columns represent attributes, features of the table. The column of the “personal data” table might be an address, a telephone number etc. Rows represent instances, actual information for each column

(attribute) presented in the table.

5.2 Details

The latest update of the CTU Database is called “brnodata_step5”. It has 65 tables, 931 columns and 2 407 425 instances in total.

These numbers are obtained from SchemaSpy software described in Chapter 3.1. Several relations among tables generated by SchemaSpy are shown in Appendix 2.

I use read mode access to the database as user “brnodata_reader”, which means that I can only read data from all tables, but I cannot modify them and I cannot change the database.

The database has 3 schemas – information_schema, public and pg_catalog. The access rights for brnodata_reader has schema “public”.

There are also 13 views. Views consist of attributes of base tables. The naming of views copy the name of the base table, but differs by ending “2”, e.g. base table is “anamneza_lekar” and its derived view is “anamneza_lekar2”. Views have sense, because they contain additional attributes with calculations of time or period of time.

Tables with “cb_” prefix are explanatory tables to attributes used in another tables. They use dials according to the data standard version 4 of Ministry of Health of the Czech Republic known as “DASTA” (source [24]).

Attribute “ic_chor” contains numbers of medical records of gravidas. It is used as database foreign key in data mining tables listed in Chapter 6.1.1. Every gravida has unique number of medical record and therefore it is possible to distinguish the data for DM.

5.3 Example of reading data from the database

To transfer data from the database to see them in a data mining software, SQL scripts (queries) are used. The example of simple SQL script is given in Figure 5.

```
SELECT ou2.ic_chor, anl.mens_od, ans.teplota FROM
osobni_udaje2 ou2 JOIN "anamneza_lekar" anl USING
(ic_chor) JOIN "anamneza_sestra" ans USING(ic_chor);
```

Figure 5: This SQL script selects attributes "ic_chor", "mens_od", "teplota" from tables "osobni_udaje2", "anamneza_lekar" and "anamneza_sestra" respectively. It uses alias names "ou2", "anl", "ans" to shorten the length in large SQL script. To connect attributes from different tables by corresponding number of medical record "ic_chor", commands "JOIN" and "USING" are necessary.

5.4 Data description and explanation

This section only describes obstetrics attributes, which are chosen for data mining. The process of selection of proper attributes is in Chapter 6.1 – Data preparation.

Attributes are listed and explained in Table 2 and Table 3. Table 2 shows annotations, it means attributes with information I would like to investigate. Table 3 shows selected attributes.

Tables include classes and number of missing instances of each attribute. The research is lead in Czech language, therefore naming of classes is not in English. I explain only the most frequent classes, e.g. classes “ANO“ or “ano” are translated as “yes“, classes “NE“ or “ne“ as “no“, classes “vetsi...” mean “higher than ...“, classes “mensi...” mean “lower than ...“.

In dependency on what I want to discover, relevant annotation from Table 2 and relevant attributes from Table 3 are chosen. Surely, for the annotation “ANOTACE_ph_7_15“ attributes “ph_7_0_int“, “ph_7_15_int“, “ph_7_20_int“ are disregarded, because they come out of the same data.

Full dataset contains 12 annotations, 54 attributes and 43 203 instances. Annotations have missing values, thus corresponding datasets contain less instances.

Table 2: Annotations of obstetrics data. (Classes and annotations are in Czech language; “NE“ means “no“, “ANO“ is “yes“, “vetsi...” means “higher than“ and “mensi...” is “lower than“)

Annotation	Explanation of annotation	Class (Number of instances)	Missing values
ANOTACE_apgar_suma_5	Evaluation of apgar sum in 5th minute	vetsi7 (42290), mensi7 (633)	280
ANOTACE_ctg_hypoxie_spatny	Suspect monitoring and hypoxia indication within childbirth	NE (12126), ANO (9218)	21 859
ANOTACE_ctg_spatny	Suspect and pathological monitoring within childbirth	NE (12422), ANO (8601)	22 180
ANOTACE_dystokie_ramen	Evaluation of dystocia of shoulder of a newborn	NE (43056), ANO (60)	87
ANOTACE_hmot_nov	Weight of fetus (grams)	2500_4000 (36464), vetsi4000 (3805), 1000_1500 (340), 1500_2500 (2334), mensi1000 (249)	11
ANOTACE_hypoxie	Hypoxia of fetus	NE (42070), ANO (1046)	87
ANOTACE_inhalace	Inhalation of fetus in delivery room	NE (35637), ANO (1399)	6 167
ANOTACE_ph_7_0	pH of newborn (division border pH = 7.0)	vetsi7 (13637), mensi7 (408)	29 158
ANOTACE_ph_7_15	pH of newborn (division border pH = 7.15)	vetsi7_15 (11450), mensi7_15 (2595)	29 158
ANOTACE_ph_7_20	pH of newborn (division borders pH = 7.0, pH = 7.20)	7_720 (4209), vetsi7_20 (9428), mensi7 (408)	29 158
ANOTACE_zpusob_porodu	Evaluation of method of delivery	SpontVag (34661), SC (7415), OperVag (1113)	14
ANOTACE_zpusob_porodu_opervag	Surgery vaginal delivery	Forceps (953), Vex (139), Forceps+Vex (21)	42 090

Table 3: Selected attributes of obstetrics data. (Classes and attributes are in Czech language; “ne“ means “no“, “ano“ is “yes“, “vetsi...” means “higher than“ and “mensi...” is “lower than“)

Attribute	Explanation of attribute	Class (Number of instances)	Missing values
anamn_abuzus_kour_corr	Anamnesis: excessive smoking	_ne (33537), _kurak_10+ (1829), _stopkurak (1184), _kurak_10- (455)	6 198
anamn_leky_corr_hypertense	Anamnesis: hypertension	ne (41151), ano (1568)	484
anamn_operace_corr_sc	Anamnesis: Caesarean section in the past	ne (40530), ano (2455)	218
apgar_suma_5	Apgar sum in 5th minute	10 (26438), 0 (278), 8 (2783), 9 (12268), 5 (56), 7 (801), 6 (251), 4 (22), 1 (13), 3 (9), 2 (4)	280
bmi_nyni_int	Body mass index of gravida, now (childbirth)	nadvaha (20593), obezita_1st (7966), norma (12105), obezita_2st (1945), podvaha (37), obezita_3st (557)	0

Attribute	Explanation of attribute	Class (Number of instances)	Missing values
bmi_pred_int	Body mass index of gravida, before pregnancy	norma (30586), nadvaha (6887), podvaha (2799), obezita_1st (2164), obezita_2st (565), obezita_3st (202)	0
cervix_skore_suma	Sum of cervix score	8 (6330), 9 (13917), 6 (4841), 4 (2557), 5 (2974), 7 (5741), 2 (1034), 10 (2767), 3 (2108), 1 (408), 0 (221)	305
doba1trv_epoch_int	Duration of 1st stage of labour (time from 1st to 2nd stage of labour)	mensilh (40531), 0h (1564), 8h_24h (11), vetsi48h (5), 1h_8h (5), 24h_48h (1)	1 086
doba2trv_epoch_int	Duration of 2nd stage of labour (time from 2nd stage of labour to delivery)	0_30min (42339), mensi0 (16), 2h_vic (4), 30min_1h (3), 1h_1h30 (1)	840
dr_vedl_doba_1_corr_analgetika	Doctor applied analgesics within 1st stage of labour	ne (33613), ano (4822)	4 768
dr_vedl_doba_1_corr_antibiotika	Doctor applied antibiotics within 1st stage of labour	ano (6877), ne (31558)	4 768
dr_vedl_doba_1_corr_epidural	Doctor applied epidural within 1st stage of labour	ne (30448), ano (7987)	4 768
dr_vedl_doba_1_corr_indukce	Doctor applied induction within 1st stage of labour	ano (10803), ne (27632)	4 768
dr_vedl_doba_1_corr_spasmolytika	Doctor applied spasmolytics within 1st stage of labour	ne (25499), ano (12936)	4 768
dr_vedl_doba_1_corr_susp_monitor	Suspect finding by CTG within 1st stage of labour	ne (36219), ano (2216)	4 768
dr_vedl_doba_1_corr_zkalena_voda	Turbid amniotic fluid within 1st stage of labour	ne (37462), ano (973)	4 768
dr_vedl_doba_2_corr_decelerace	Doctor registered deceleration within 2nd stage of labour	ne (40349), ano (2767)	87
dr_vedl_doba_2_corr_diabetes	Diabetes present	ne (42882), ano (234)	87
dr_vedl_doba_2_corr_dystokie_ramen	Doctor registered dystocia of shoulder within 2nd stage of labour	ne (43056), ano (60)	87
dr_vedl_doba_2_corr_forceps	Doctor applied forceps within 2nd stage of labour	ne (42314), ano (802)	87
dr_vedl_doba_2_corr_hypoxie	Doctor registered hypoxia of fetus within 2nd stage of labour	ne (42070), ano (1046)	87
dr_vedl_doba_2_corr_iugr	Doctor registered IUGR within 2nd stage of labour	ne (43091), ano (25)	87
dr_vedl_doba_2_corr_nepostupujiciporod	Doctor registered no move in delivery within 2nd stage of labour	ne (42304), ano (812)	87
dr_vedl_doba_2_corr_oxytocin	Doctor applied oxytocin within 2nd stage of labour	ne (33183), ano (9933)	87
dr_vedl_doba_2_corr_preeklampsie	Doctor registered preeclampsia of gravida within 2nd stage of labour	ne (43037), ano (79)	87
dr_vedl_doba_2_corr_primarnisc	Primary Caesarean section	ne (42996), ano (120)	87
dr_vedl_doba_2_corr_pupecnikkk	Umbilical cord around neck within 2nd stage of labour	ne (42625), ano (491)	87
dr_vedl_doba_2_corr_sc	Caesarean section within 2nd stage of labour	ne (37441), ano (5675)	87
dr_vedl_doba_2_corr_spontzahl	Spontaneous delivery within 2nd stage of labour	ne (42080), ano (1036)	87

Attribute	Explanation of attribute	Class (Number of instances)	Missing values
hm_placenta_int	Weight of placenta after the delivery	600_1000 (11431), 400_600 (29689), 250_400 (1034), vetsi1000 (42), mensi250 (138)	869
hm_rozdil_int	Weight difference of gravida (childbirth and before pregnancy)	16_25 (9035), 8_16 (29045), 0_8 (4405), mensi0 (201), vetsi25 (517)	0
hmot_nov_int	Weight of fetus in grams	2500_4000 (36464), vetsi4000 (3805), 1000_1500 (340), 1500_2500 (2334), mensi1000 (249)	11
kod_pohl_nov_vys	Sex of newborn	chlapec (22208), devce (20962)	33
krev_ztrata_int	Blood loss within childbirth	1_500 (41152), 500_1000 (1485), 1000vic (143), 0 (1)	422
lecba_sal_1_vys	Medical treatment of newborn, hospital hall 1	zadna (33535), kyslik (3778), UPV_maska (726), intubace (293), leky (7), masaz_srdce (9)	4 855
ph_7_0_int	pH of newborn (division border pH = 7.0)	vetsi7 (13637), mensi7 (408)	29 158
ph_7_15_int	pH of newborn (division border pH = 7.15)	vetsi7_15 (11450), mensi7_15 (2595)	29 158
ph_7_20_int	pH of newborn (division borders pH = 7.0, pH = 7.20)	7_720 (4209), vetsi7_20 (9428), mensi7 (408)	29 158
plod_voda_kod_vys	Quality of amniotic fluid	cista (39346), zkalena (3490)	367
poloha_plodu_kod_vys	Position of fetus within childbirth	zhlavim (38452), koncem_panevnim (2205), jina (1855)	691
porod_kleste_kod_vys	Forceps used within childbirth	ne (42330), ano (872)	1
porod_vex_kod_vys	Vacuum extractor used within childbirth	ne (43035), ano (167)	111
porody_celk	Total number of deliveries of gravida (prior to current delivery)	1 (16400), 0 (21339), 2 (3976), 4 (306), 3 (883), 12 (1), 6 (42), 5 (97), 8 (15), 7 (26), 10 (3), 9 (3), 13 (1)	111
pozn_corr_pupecnikkk	Umbilical cord around neck	ne (29612), ano (7424)	6 167
pozn_corr_sag	SAG	ne (31117), ano (5919)	6 167
shrnuti_terapie_corr	Summarization of medical treatment applied within childbirth	MEM (11997), Oxytocin (8751), Jina (278), Prostin (4520)	17 657
teplota_int	Temperature of gravida (degree Celsius)	mensi365 (3704), 365_37 (38811), 375_38 (138), 37_375 (484), vetsi38 (66)	0
tk_dia_int	Diastolic blood pressure of gravida	optim (23363), hypertenze (4042), normal (10543), vysoky (5255)	0
tk_syst_int	Systolic blood pressure of gravida	optim (12486), hypertenze (4201), normal (17217), vysoky (9299)	0
tt_dokonceny36_int	Pregnancy week (1st interval)	vetsi36 (41119), 28_36 (1909), mensi28 (175)	0
vek_matky_int	Age of gravida (time of childbirth)	17_36 (40506), vetsi36 (2582), mensi17 (85)	30
vysetr_ctg_corr	Examination of gravida by cardiotocography	Fyziologický (15780), Suspektní (2717), Jiný (2058), Patologický (80)	22 568
vysetr_uzv_poloha_corr	Position of fetus, ultrasonic examination	PPKP (531), PPHI (4241), Jina (2319), PPHI (1527)	34 585
způsob_por_kod_vys	Method of delivery (division only to vaginal and Caesarean)	vaginálne (35774), SC (7415)	14

6 Experiments

In this chapter I present experimental results.

In section 6.1 is data preparation of HIS data. I create datasets, which I explore via association rules in Chapter 6.2 by software Lisp-Miner. I present attributes or combination of attributes that imply significant situations before or within childbirth, which could serve as an aid or new knowledge for doctors of FN Brno hospital. The most important results are explained and interpreted in Chapter 6.3.

In Appendix 3 is an example howto create dataset in RapidMiner software.

6.1 Data preparation

I make following steps in order to prepare proper datasets for purposes of Biodat Research Group from HIS:

- Manual selection of attributes from CTU Database (Chapter 6.1.1),
- Definition of limitations of attributes (Chapter 6.1.2),
- Creation of new attributes, categorization, adjusting names of attributes and classes (Chapter 6.1.3),
- Creation and completion of datasets according to selection of attributes by doctors of FN Brno hospital (Chapter 6.1.4).

6.1.1 Manual selection of attributes

Because of enormous data size, in cooperation with my supervisor, we had to agree on areas, which could be interesting or which could influence normal or pathological states before or within delivery of singleton newborns. Areas are “pH measure“ (potential of hydrogen), “apgar sum in 5th minute“, “weight of fetus“, “method of childbirth“, “hypoxia of fetus“, “suspect monitoring of fetus“ (CTG - cardiotocography), “inhalation of fetus in delivery room“ and “dystocia of shoulder of a newborn“.

Based on next steps and chapters, defined areas became annotations (main attributes of datasets), which are listed in Table 2 of Chapter 5.4.

CTU Database contains almost 500 possible attributes to do a research in the defined areas. Manual selection through possible attributes resulted in set of 106 attributes (118 with annotations). Tables of CTU Database 5th version (step5) are “jpt_all2“ (significant attributes from other tables are joined here), “anamneza_lekar2“ (anamnesis from doctors view), “pp_objekt_nalez2“ (medical examination), “anamneza_sestra“ (anamnesis from nurses view) and “porod_shrnuti“ (summary of delivery).

Manual selection in this case means:

- Reduction of multiple attributes, which relate to the same description (e.g. attributes of multiple summary - “shrnuti_shrnuti“, “shrnuti_terapie“, “shrnuti_doktori“ and only corrected summary “shrnuti_terapie_corr“ is chosen),
- Removal of attributes that evidently have no impact on defined areas (e.g. address, record number or number of insurance company),
- Removal of attributes that should not have impact on defined areas (e.g. allergy of father or different anamneses of gravida).
- Elimination of redundant attributes (e.g. attribute “anamn_transfuzne_corr_ne“ is empty) or of highly imbalanced attributes except “dystocia of shoulder of a newborn“ (e.g. attribute “anamn_otec_alergie_corr_jineleky“ has class “true“ with almost 30 000 instances and class “false“ with only 35 instances),

I would like to note that a mistake can be made here, because the selection is not scientifically supported (for example by entropy or odd ratios).

6.1.2 Physiological limitations

Several attributes could contain errors and are outside physiological range. Limitations or intervals of attributes are defined according to Table 4 to minimize errors of results.

These limits shrink CTU Database by approximately 4 000 instances (from more than 47 000 to more than 43 000 instances) because limits of attributes have to be valid simultaneously.

Table 4: Limitations of attributes that could contain errors.

Attribute	Explanation of attribute	Limit, Range
ph	pH of newborn	between 5 and 10
tk_syst	Systolic blood pressure of gravida	between 60 and 160
tk_dia	Diastolic blood pressure of gravida	between 40 and 120
pan_d_bisp	Pelvis proportion of gravida – distantia bispinalis	between 15 and 45
ans.pan_d_bicri	Pelvis proportion of gravida – distantia bicrinalis	between 15 and 45
pan_d_bitr	Pelvis proportion of gravida – distantia bitrochanterica	between 15 and 45
pan_c_ext	Pelvis proportion of gravida – conjugata externa	between 15 and 45
tep_min	Minimum heart beat of gravida	between 30 and 200
teplota	Temperature of gravida	between 35.5 and 42
vyska	Height of gravida	between 140 and 210
hm_pred	Weight of gravida, before pregnancy	between 40 and 150
hm_nyni	Weight of gravida, now (childbirth)	between 40 and 170
tt_dokoceny	Pregnancy week	between 20 and 50

6.1.3 New attributes, Categorization, Adjusting names of classes and attributes

In broader context, new attributes are all annotations (see Table 2) and attributes with ending “_int” and “_vys” (see Table 3) because they are derived from attributes presented in the original CTU Database.

To complete information in this section below (e.g. conditions of attributes), it is good to watch materials in CD, which is enclosed to this diploma thesis.

In this section, I consider **new attributes** as attributes calculated from more than one attribute of CTU Database. Therefore new attributes are “bmi_nyni_int” (current BMI of gravida), “bmi_pred_int” (BMI of gravida before pregnancy) and “hm_rozdil_int” (weight difference of gravida). BMI is calculated from weight of gravida, which is divided by square of height of gravida. Weight difference is how many kilograms gravida gain within pregnancy, thus it is calculated as subtraction of weight of gravida before pregnancy from weight of gravida before delivery.

I categorize (create classes of) attributes and annotations to reduce complexity of data and to spare time within search for association rules.

Categorization of attributes has ending “_int” (see Table 3). BMI is categorized according to [22] and blood pressure according to [23]. Duration of stages of labour (doba_1_trv_epoch_int, doba2_trv_epoch_int) is categorized after transformation from milliseconds to hours (doba_1_trv_epoch_int) or to hours and minutes (doba_2_trv_epoch_int). Pregnancy week of delivery (attribute tt_dokonceny36_int) has 3 categories, e.g. category “mensi28” states that newborn is delivered before 28th pregnancy week, category “28_36” states delivery between 28th and 36th pregnancy week and category “vetsi36” states delivery after 36th pregnancy week. The rest of attributes with ending “_int” is self-explanatory.

Categorization of annotations copy categorization of attributes except cases of suspect monitoring and delivery, to be specific of annotations:

- ANOTACE_ctg_spatny, ANOTACE_ctg_hypoxie_spatny,
- ANOTACE_zpusob_porodu, ANOTACE_zpusob_porodu_opervag.

“ANOTACE_ctg_spatny” means suspect monitoring within childbirth, thus category “NE” (no) is positive outcome and category “ANO” (yes) is negative outcome. Positive outcome is defined as physiological state of medical examination of gravida, no deceleration within 2nd stage of labour and no suspect monitoring of cardiotocography within 1st stage of labour. Negative outcome is the opposite, thus there is deceleration or suspect monitoring or pathological state of medical examination of gravida.

Annotation “ANOTACE_ctg_hypoxie_spatny” is similar to annotation “ANOTACE_ctg_spatny”, but in addition it covers states of hypoxia of fetus within 2nd stage of labour.

Annotation “ANOTACE_zpusob_porodu” means method of delivery. It differs from attribute of method of delivery that it separates vaginal childbirth to spontaneous (category “SpontVag”) and operational (category “OperVag”).

Annotation “ANOTACE_zpusob_porodu_opervag” focus on operational method of childbirth. It separates this method to applied tools, i.e. forceps used (category “Forceps”), vacuum extraction of newborn (category “Vex”) and both used together (category “Forceps+Vex”).

I have changed **names of certain attributes and classes** to be more explanatory. For example, sex of newborn was attribute “kod_pohl” and has classes or categories “1”

and “2“. First problem could be to get to know that this attribute relates to newborn, the second could be to find out that class “1“ is male and class “2“ female. Resulting attribute is called “kod_pohl_nov_vys“ and has classes “chlapec“ (male) and “děvče“ (female).

Ending “_vys“ in the mentioned attribute and the rest of attributes of Table 3 states that classes are transformed to natural Czech language.

All changes above are done via SQL scripts while reading data from CTU Database.

Regarding boolean attributes (true, false; in CTU Database “t“ and “f“ respectively), I use “Map“ operator in RadidMiner to rename “t“ to “ano“ (yes) and “f“ to “ne“ (no). Again, the explanation of resulting association rules is much easier.

6.1.4 Datasets creation and completion

To complete the data preparation step, set of chosen 106 attributes (see Chapter 6.1.1) was given to doctors at FN Brno hospital. Doctors defined 54 attributes that are significant for them or that might have influence on annotations. List of these attributes is in Table 3.

Afterwards, I create datasets in RadidMiner software according to Figure 6 and Figure 7.

Figure 6 presents process of creating 2 entire datasets into Weka file “.arff“. One entire dataset comprises annotations and attributes. The second entire dataset has also annotations and attributes, but there is one additional attribute “ic_chor“, which is used as unique identifier of the dataset and in “Lisp-Miner“ software as a primary key.

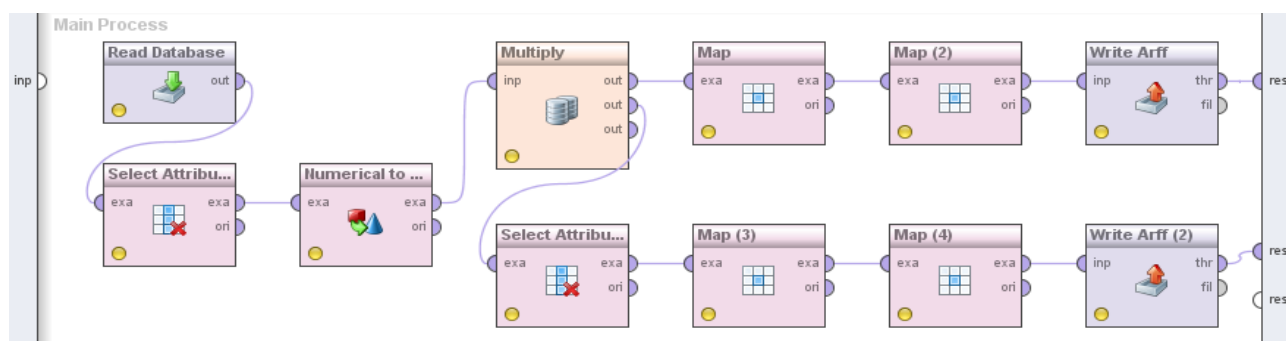


Figure 6: RapidMiner process, which creates entire datasets into Weka files from CTU Database.

This process reads data from CTU Database, selects proper attributes and annotations and changes format of data (numerical data becomes classes) in “Numerical to Polynomial“

operator. Operator “Multiply“ only switch process to two branches. The upper branch creates entire dataset without attribute “ic_chor“, the bottom branch with attribute “ic_chor“. “Map“ or “Map (3)“ operators convert boolean value true to “ano“ value and “Map (2)“ and “Map (4)“ operators convert boolean value false to “ne“ value (as described in the end of Chapter 6.1.3).

Figure 7 presents process of creating 3 datasets of annotation “ANOTACE_hmot_nov“. Two datasets are Weka files without attribute “ic_chor“ and one dataset is Microsoft Access Database file “.mdb“ with attribute “ic_chor“.

Every annotation uses special RapidMiner process.

The process starts by reading entire dataset, then it selects proper attributes for annotation and label the annotation as main attribute (“Set Role“ operator). To remove missing values from annotation, “Filter Examples“ operator is used. Operator “Multiply“ switch the process to 3 branches. The upper branch after “Multiply“ operator represents creating nominal dataset (classes of attributes) into Weka file, the middle branch represents creating numerical dataset (“Nominal to Numerical“ operator) – necessary for Orange 2.0 software and the bottom branch is the creation of Microsoft Access Database file (“Write Access“ operator; used by Lisp-Miner software).

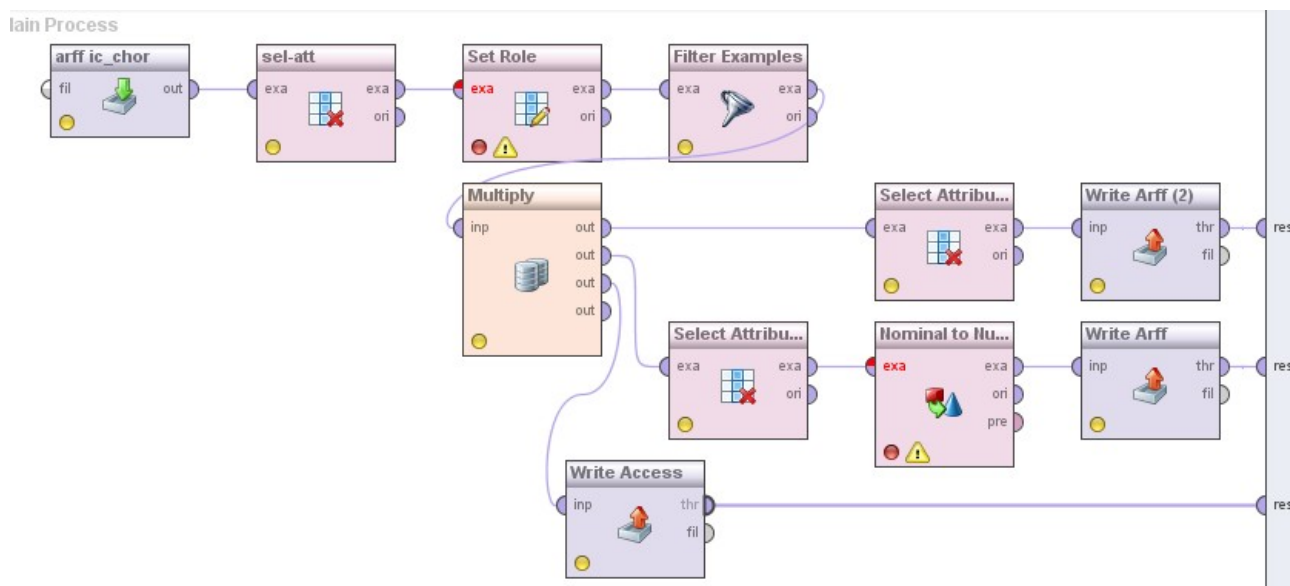


Figure 7: RapidMiner process, which creates datasets of weight of newborn into Weka files and Microsoft Access Database file from entire dataset. Operator “Read ARFF“ is renamed to “arff ic_chor“ and operator “Select Attributes“ to “sel-att“.

6.2 Results

In cooperation with my supervisor, we defined 4 questions, which are good representation of questions that the data should be able to answer:

- **Question 1:** Will the outcome of the newborn be normal?
- **Question 2:** Which factors within labour result in or influence most delivery by Caesarean?
- **Question 3:** Is there any relation between number of previous deliveries and the outcome of current (ongoing) labour?
- **Question 4:** Which factors in anamnesis of gravida relate to macrosomia of newborns (weight of fetus higher than 4 000 grams)?

Important states should come out of association rules of datasets, which can answer these questions. Good outcome of delivery describes pH and apgar sum in 5th minute. Thus to answer Question 1, I choose datasets with annotations “ANOTACE_apgar_suma_5” and “ANOTACE_ph_7_15”. Question 2 and Question 3 relate to evaluation of childbirth, therefore I choose dataset with annotation “ANOTACE_zpusob_porodu”. To answer Question 4, I choose dataset with annotation “ANOTACE_hmot_nov”, because this question relates to weight of fetus.

To obtain results of association rules, I use Lisp-Miner software and datasets with above mentioned annotations in Microsoft Access Database files created as in Figure 7.

Datasets for Lisp-Miner are set according to “Tutorials” section of Lisp-Miner official website (source [18]). I use possibility “Each value – one category” to keep attributes without changes (no aggregation of classes) as defined in Table 3. Then I set Lisp-Miner task to generate results. Firstly I use only one attribute as antecedent and one class of proper annotation as succedent. Secondly, interesting attributes are then used in combinations to specify more interesting association rules.

Parameters of Lisp-Miner task are quantifier, its settings and base value (minimum number of instances that holds for antecedent imply succedent) or ceiling value (maximum number of instances).

For my experiments, ceiling value is useless, thus I use base value.

According to me, the best quantifier is Fisher quantifier, which performs one-sided Fisher test. Fisher quantifier is fast and proper enough to go through the entire dataset of an annotation. It works in a manner that it rejects independence of antecedent and succedent on level α . In other words, it confirms the dependence on level $(1 - \alpha)$. For instance when α is 0.005, there is 99.5% chance that there is dependency between antecedent and succedent.

Settings of quantifier, base value, antecedent and succedent are in next chapters that deal with defined questions. I demonstrate used method in Chapter 6.2.1 for annotation “ANOTACE_apgar_suma_5” (Figure 8, Figure 9). I do not present results of antecedent of length from 1 to 1 in the rest of annotations (for this, see Lisp-Miner tasks on CD enclosed to this diploma thesis). Thus I only provide the final result (Figure 10 to Figure 13).

6.2.1 Association rules, Question 1

The positive outcome of delivery is given by apgar sum in 5th minute with value higher than 7 and by pH value higher than 7.15.

Thus first setup of Lisp-Miner task is Fisher quantifier with base 100 instances and α value 0.001. Antecedent has length from 1 to 1, therefore antecedent can be one of classes of all dataset. Succedent is class “vetsi7” of annotation “ANOTACE_apgar_suma_5”. Results of this task are shown in Figure 8. Interesting classes mark blue color in Figure 8.

Second setup of class “vetsi 7” of annotation “ANOTACE_apgar_suma_5” has antecedent length from 1 to 6 (from 1 up to 6 classes in conjunction). Antecedents are selected classes from Figure 8 (blue color). The rest of setup is the same as in previous case. Results are shown in Figure 9.

I apply the same steps for annotation “ANOTACE_ph_7_15” as well. I present results of antecedent length from 1 to 6 in Figure 10. Task setup remains the same. However, I focus on negative outcome (class “mensi7_15”, pH value lower than 7.15). Here is the assumption that discovered negative outcome also represents positive outcome of childbirth in remaining classes of attributes or association rules.

Actual group of hypotheses: All hypothesis

Hypotheses in group: 28 Shown hypotheses: 28 Highlighted: 0

Nr.	Id	Conf	Hypothesis
1	17	0.997	Lecba_sal_1_vys(zadna) *** ANOTACE_apgar_suma_5(vetsi7)
2	4	0.993	Cervix_skore_suma(9) *** ANOTACE_apgar_suma_5(vetsi7)
3	26	0.992	Pozn_corr_sag(ano) *** ANOTACE_apgar_suma_5(vetsi7)
4	3	0.991	Cervix_skore_suma(8) *** ANOTACE_apgar_suma_5(vetsi7)
5	14	0.990	Hmot_nov_int(2500_4000) *** ANOTACE_apgar_suma_5(vetsi7)
6	7	0.990	Dr_vedl_doba_1_corr_spasmolytika(ano) *** ANOTACE_apgar_suma_5(vetsi7)
7	27	0.990	Tt_dokonceny36_int(vetsi36) *** ANOTACE_apgar_suma_5(vetsi7)
8	10	0.989	Dr_vedl_doba_2_corr_oxytocin(ano) *** ANOTACE_apgar_suma_5(vetsi7)
9	25	0.989	Pozn_corr_pupecnikkk(ano) *** ANOTACE_apgar_suma_5(vetsi7)
10	28	0.988	Zpusob_por_kod_vys(vaginalne) *** ANOTACE_apgar_suma_5(vetsi7)
11	24	0.988	Porody_celk(1) *** ANOTACE_apgar_suma_5(vetsi7)
12	12	0.988	Hm_placenta_int(400_600) *** ANOTACE_apgar_suma_5(vetsi7)
13	11	0.987	Dr_vedl_doba_2_corr_sc(ne) *** ANOTACE_apgar_suma_5(vetsi7)
14	13	0.987	Hm_rozdil_int(8_16) *** ANOTACE_apgar_suma_5(vetsi7)
15	15	0.987	Kod_pohl_nov_vys(devce) *** ANOTACE_apgar_suma_5(vetsi7)
16	2	0.987	Bmi_nyri_int(nadvaha) *** ANOTACE_apgar_suma_5(vetsi7)
17	5	0.987	Doba_ltrv_epoch_int(mensi1h) *** ANOTACE_apgar_suma_5(vetsi7)
18	22	0.987	Poloha_plodu_kod_vys(zahlavim) *** ANOTACE_apgar_suma_5(vetsi7)
19	6	0.986	Dr_vedl_doba_1_corr_antibiotika(ne) *** ANOTACE_apgar_suma_5(vetsi7)
20	21	0.986	Plod_voda_kod_vys(cista) *** ANOTACE_apgar_suma_5(vetsi7)
21	16	0.986	Krev_ztrata_int(1_500) *** ANOTACE_apgar_suma_5(vetsi7)
22	9	0.986	Dr_vedl_doba_2_corr_hypoxie(ne) *** ANOTACE_apgar_suma_5(vetsi7)
23	1	0.986	Anamn_leky_corr_hypertense(ne) *** ANOTACE_apgar_suma_5(vetsi7)
24	23	0.986	Porod_kleste_kod_vys(ne) *** ANOTACE_apgar_suma_5(vetsi7)
25	8	0.985	Dr_vedl_doba_2_corr_dystokie_ramen(ne) *** ANOTACE_apgar_suma_5(vetsi7)
26	20	0.976	Ph_7_20_int(vetsi7_20) *** ANOTACE_apgar_suma_5(vetsi7)
27	19	0.974	Ph_7_15_int(vetsi7_15) *** ANOTACE_apgar_suma_5(vetsi7)
28	18	0.969	Ph_7_0_int(vetsi7) *** ANOTACE_apgar_suma_5(vetsi7)

Figure 8: Results of Lisp-Miner task with Fisher quantifier, antecedent of length from 1 to 1 and succedent is class “vetsi7“ of apgar sum in 5th minute.

Actual group of hypotheses: All hypothesis

Hypotheses in group: 23 Shown hypotheses: 23 Highlighted: 0

Nr.	Id	Conf	Hypothesis
1	10	1.000	Dr_vedl_doba_2_corr_oxytocin(ano) & Porody_celk(1) & Pozn_corr_sag(ano) & Tt_dokonceny36_int(vetsi36) *** ANOTACE_apgar_suma_5(vetsi7)
2	19	0.998	Porody_celk(1) & Pozn_corr_sag(ano) & Tt_dokonceny36_int(vetsi36) *** ANOTACE_apgar_suma_5(vetsi7)
3	13	0.997	Dr_vedl_doba_2_corr_oxytocin(ano) & Pozn_corr_sag(ano) & Tt_dokonceny36_int(vetsi36) *** ANOTACE_apgar_suma_5(vetsi7)
4	11	0.995	Dr_vedl_doba_2_corr_oxytocin(ano) & Porody_celk(1) & Tt_dokonceny36_int(vetsi36) *** ANOTACE_apgar_suma_5(vetsi7)
5	6	0.995	Dr_vedl_doba_1_corr_spasmolytika(ano) & Pozn_corr_sag(ano) & Tt_dokonceny36_int(vetsi36) *** ANOTACE_apgar_suma_5(vetsi7)
6	4	0.995	Dr_vedl_doba_1_corr_spasmolytika(ano) & Porody_celk(1) & Tt_dokonceny36_int(vetsi36) *** ANOTACE_apgar_suma_5(vetsi7)
7	22	0.995	Pozn_corr_sag(ano) & Tt_dokonceny36_int(vetsi36) *** ANOTACE_apgar_suma_5(vetsi7)
8	18	0.995	Porody_celk(1) & Pozn_corr_sag(ano) *** ANOTACE_apgar_suma_5(vetsi7)
9	12	0.995	Dr_vedl_doba_2_corr_oxytocin(ano) & Pozn_corr_sag(ano) *** ANOTACE_apgar_suma_5(vetsi7)
10	5	0.994	Dr_vedl_doba_1_corr_spasmolytika(ano) & Pozn_corr_sag(ano) *** ANOTACE_apgar_suma_5(vetsi7)
11	3	0.994	Dr_vedl_doba_1_corr_spasmolytika(ano) & Porody_celk(1) *** ANOTACE_apgar_suma_5(vetsi7)
12	9	0.994	Dr_vedl_doba_2_corr_oxytocin(ano) & Porody_celk(1) *** ANOTACE_apgar_suma_5(vetsi7)
13	15	0.993	Porody_celk(0) & Pozn_corr_sag(ano) & Tt_dokonceny36_int(vetsi36) *** ANOTACE_apgar_suma_5(vetsi7)
14	20	0.992	Porody_celk(1) & Tt_dokonceny36_int(vetsi36) *** ANOTACE_apgar_suma_5(vetsi7)
15	21	0.992	Pozn_corr_sag(ano) *** ANOTACE_apgar_suma_5(vetsi7)
16	7	0.992	Dr_vedl_doba_1_corr_spasmolytika(ano) & Tt_dokonceny36_int(vetsi36) *** ANOTACE_apgar_suma_5(vetsi7)
17	14	0.991	Dr_vedl_doba_2_corr_oxytocin(ano) & Tt_dokonceny36_int(vetsi36) *** ANOTACE_apgar_suma_5(vetsi7)
18	1	0.990	Dr_vedl_doba_1_corr_spasmolytika(ano) *** ANOTACE_apgar_suma_5(vetsi7)
19	23	0.990	Tt_dokonceny36_int(vetsi36) *** ANOTACE_apgar_suma_5(vetsi7)
20	2	0.990	Dr_vedl_doba_1_corr_spasmolytika(ano) & Porody_celk(0) & Tt_dokonceny36_int(vetsi36) *** ANOTACE_apgar_suma_5(vetsi7)
21	8	0.989	Dr_vedl_doba_2_corr_oxytocin(ano) *** ANOTACE_apgar_suma_5(vetsi7)
22	17	0.988	Porody_celk(1) *** ANOTACE_apgar_suma_5(vetsi7)
23	16	0.988	Porody_celk(0) & Tt_dokonceny36_int(vetsi36) *** ANOTACE_apgar_suma_5(vetsi7)

Figure 9: Results of Lisp-Miner task with Fisher quantifier, antecedent is selected attributes from Figure 8 of length from 1 to 6 and succedent is class “vetsi7“ of apgar sum in 5th minute.

Actual group of hypotheses: All hypothesis

Hypotheses in group: 437 Shown hypotheses: 437 Highlighted: 0

Nr.	Id	Conf	Hypothesis
1	373	0.565	Dr_vedl_doba_2_corr_decelerate(ano) & Dr_vedl_doba_2_corr_forceps(ano) *** ANOTACE_ph_7_15(mensi7_15)
2	375	0.563	Dr_vedl_doba_2_corr_decelerate(ano) & Dr_vedl_doba_2_corr_forceps(ano) & Poloha_plodu_kod_vys(zahlavim) *** ANOTACE_ph_7_15(mensi7_15)
3	377	0.560	Dr_vedl_doba_2_corr_decelerate(ano) & Dr_vedl_doba_2_corr_forceps(ano) & Porody_celk(0) *** ANOTACE_ph_7_15(mensi7_15)
4	376	0.554	Dr_vedl_doba_2_corr_decelerate(ano) & Dr_vedl_doba_2_corr_forceps(ano) & Poloha_plodu_kod_vys(zahlavim) & Porody_celk(0) *** ANOTACE_ph_7_15(mensi7_15)
5	178	0.554	Dr_vedl_doba_1_corr_epidural(ano) & Dr_vedl_doba_1_corr_indukce(ano) & Dr_vedl_doba_2_corr_decelerate(ano) & Porody_celk(0) *** ANOTACE_ph_7_15(mensi7_15)
6	177	0.553	Dr_vedl_doba_1_corr_epidural(ano) & Dr_vedl_doba_1_corr_indukce(ano) & Dr_vedl_doba_2_corr_decelerate(ano) & Poloha_plodu_kod_vys(zahlavim) & Porody_celk(0) *** AN
7	199	0.544	Dr_vedl_doba_1_corr_epidural(ano) & Dr_vedl_doba_1_corr_spasmolytika(ano) & Dr_vedl_doba_2_corr_decelerate(ano) & Poloha_plodu_kod_vys(zahlavim) & Porody_celk(0) **
8	200	0.543	Dr_vedl_doba_1_corr_epidural(ano) & Dr_vedl_doba_1_corr_spasmolytika(ano) & Dr_vedl_doba_2_corr_decelerate(ano) & Porody_celk(0) *** ANOTACE_ph_7_15(mensi7_15)
9	197	0.537	Dr_vedl_doba_1_corr_epidural(ano) & Dr_vedl_doba_1_corr_spasmolytika(ano) & Dr_vedl_doba_2_corr_decelerate(ano) & Hm_rozdl_int(8_16) *** ANOTACE_ph_7_15(mensi7_15)
10	374	0.534	Dr_vedl_doba_2_corr_decelerate(ano) & Dr_vedl_doba_2_corr_forceps(ano) & Hm_rozdl_int(8_16) *** ANOTACE_ph_7_15(mensi7_15)
11	174	0.534	Dr_vedl_doba_1_corr_epidural(ano) & Dr_vedl_doba_1_corr_indukce(ano) & Dr_vedl_doba_2_corr_decelerate(ano) *** ANOTACE_ph_7_15(mensi7_15)
12	330	0.527	Dr_vedl_doba_1_corr_spasmolytika(ano) & Dr_vedl_doba_2_corr_forceps(ano) & Poloha_plodu_kod_vys(zahlavim) & Porody_celk(0) *** ANOTACE_ph_7_15(mensi7_15)
13	329	0.525	Dr_vedl_doba_1_corr_spasmolytika(ano) & Dr_vedl_doba_2_corr_forceps(ano) & Poloha_plodu_kod_vys(zahlavim) *** ANOTACE_ph_7_15(mensi7_15)
14	176	0.525	Dr_vedl_doba_1_corr_epidural(ano) & Dr_vedl_doba_1_corr_indukce(ano) & Dr_vedl_doba_2_corr_decelerate(ano) & Poloha_plodu_kod_vys(zahlavim) *** ANOTACE_ph_7_15
15	175	0.524	Dr_vedl_doba_1_corr_epidural(ano) & Dr_vedl_doba_1_corr_indukce(ano) & Dr_vedl_doba_2_corr_decelerate(ano) & Hm_rozdl_int(8_16) *** ANOTACE_ph_7_15(mensi7_15)
16	196	0.519	Dr_vedl_doba_1_corr_epidural(ano) & Dr_vedl_doba_1_corr_spasmolytika(ano) & Dr_vedl_doba_2_corr_decelerate(ano) *** ANOTACE_ph_7_15(mensi7_15)
17	198	0.516	Dr_vedl_doba_1_corr_epidural(ano) & Dr_vedl_doba_1_corr_spasmolytika(ano) & Dr_vedl_doba_2_corr_decelerate(ano) & Poloha_plodu_kod_vys(zahlavim) *** ANOTACE_ph_
18	227	0.512	Dr_vedl_doba_1_corr_epidural(ano) & Dr_vedl_doba_2_corr_forceps(ano) & Poloha_plodu_kod_vys(zahlavim) *** ANOTACE_ph_7_15(mensi7_15)
19	331	0.512	Dr_vedl_doba_1_corr_spasmolytika(ano) & Dr_vedl_doba_2_corr_forceps(ano) & Porody_celk(0) *** ANOTACE_ph_7_15(mensi7_15)
20	327	0.511	Dr_vedl_doba_1_corr_spasmolytika(ano) & Dr_vedl_doba_2_corr_forceps(ano) *** ANOTACE_ph_7_15(mensi7_15)
21	145	0.509	Dr_vedl_doba_1_corr_antibiotika(ano) & Dr_vedl_doba_2_corr_decelerate(ano) & Porody_celk(0) *** ANOTACE_ph_7_15(mensi7_15)
22	225	0.505	Dr_vedl_doba_1_corr_epidural(ano) & Dr_vedl_doba_2_corr_decelerate(ano) & Porody_celk(0) *** ANOTACE_ph_7_15(mensi7_15)
23	222	0.502	Dr_vedl_doba_1_corr_epidural(ano) & Dr_vedl_doba_2_corr_decelerate(ano) & Hm_rozdl_int(8_16) & Porody_celk(0) *** ANOTACE_ph_7_15(mensi7_15)
24	224	0.501	Dr_vedl_doba_1_corr_epidural(ano) & Dr_vedl_doba_2_corr_decelerate(ano) & Poloha_plodu_kod_vys(zahlavim) & Porody_celk(0) *** ANOTACE_ph_7_15(mensi7_15)
25	218	0.501	Dr_vedl_doba_1_corr_epidural(ano) & Dr_vedl_doba_2_corr_decelerate(ano) *** ANOTACE_ph_7_15(mensi7_15)

Figure 10: Results of Lisp-Miner task with Fisher quantifier, antecedent is selected attributes (not presented in diploma thesis) of length from 1 to 6 and succedent is class “mensi7_15” of pH of newborn.

6.2.2 Association rules, Question 2

In this section, I inspect annotation “ANOTACE_zpusob_porodu” and its class “SC” (Caesarean section).

Task setup is also Fisher quantifier with base 100 instances and alpha level 0.001. Length of antecedent is from 1 to 5. Succedent is class “SC” of the annotation.

Results present Figure 11.

6.2.3 Association rules, Question 3

This task deals with previous childbirths (attribute “Porody_celk”) and their influence on the current childbirth (annotation “ANOTACE_zpusob_porodu”).

The task is easy in this case, antecedent is only one attribute “Porody_celk” and the succedent all classes of the annotation. Due to one attribute as antecedent, it does not matter on length of the antecedent. The setup of Lisp-Miner task is again Fisher quantifier, base 50 instances and alpha level 0.001.

Results are in Figure 12.

Actual group of hypotheses: All hypothesis

Hypotheses in group: 56 Shown hypotheses: 56 Highlighted: 0

Nr.	Id	Conf	Hypothesis
1	24	0.969	Bmi_nyni_int(obezita_1st) & Dr_vedl_doba_2_corr_hypoxie(ano) *** ANOTACE_zpusob_porodu(SC)
2	25	0.966	Bmi_nyni_int(obezita_1st) & Dr_vedl_doba_2_corr_hypoxie(ano) & Poloha_plodu_kod_vys(zahlavim) *** ANOTACE_zpusob_porodu(SC)
3	38	0.956	Bmi_pred_int(nadvaha) & Dr_vedl_doba_2_corr_hypoxie(ano) *** ANOTACE_zpusob_porodu(SC)
4	19	0.956	Bmi_nyni_int(obezita_1st) & Bmi_pred_int(nadvaha) & Dr_vedl_doba_2_corr_hypoxie(ano) *** ANOTACE_zpusob_porodu(SC)
5	20	0.950	Bmi_nyni_int(obezita_1st) & Bmi_pred_int(nadvaha) & Dr_vedl_doba_2_corr_hypoxie(ano) & Poloha_plodu_kod_vys(zahlavim) *** ANOTACE_zpusob_porodu(SC)
6	39	0.949	Bmi_pred_int(nadvaha) & Dr_vedl_doba_2_corr_hypoxie(ano) & Poloha_plodu_kod_vys(zahlavim) *** ANOTACE_zpusob_porodu(SC)
7	52	0.947	Dr_vedl_doba_2_corr_hypoxie(ano) *** ANOTACE_zpusob_porodu(SC)
8	13	0.945	Bmi_nyni_int(norma) & Dr_vedl_doba_2_corr_hypoxie(ano) *** ANOTACE_zpusob_porodu(SC)
9	9	0.944	Bmi_nyni_int(norma) & Bmi_pred_int(norma) & Dr_vedl_doba_2_corr_hypoxie(ano) *** ANOTACE_zpusob_porodu(SC)
10	53	0.944	Dr_vedl_doba_2_corr_hypoxie(ano) & Poloha_plodu_kod_vys(zahlavim) *** ANOTACE_zpusob_porodu(SC)
11	14	0.943	Bmi_nyni_int(norma) & Dr_vedl_doba_2_corr_hypoxie(ano) & Poloha_plodu_kod_vys(zahlavim) *** ANOTACE_zpusob_porodu(SC)
12	42	0.941	Bmi_pred_int(norma) & Dr_vedl_doba_2_corr_hypoxie(ano) *** ANOTACE_zpusob_porodu(SC)
13	10	0.939	Bmi_nyni_int(norma) & Bmi_pred_int(norma) & Dr_vedl_doba_2_corr_hypoxie(ano) & Poloha_plodu_kod_vys(zahlavim) *** ANOTACE_zpusob_porodu(SC)
14	43	0.937	Bmi_pred_int(norma) & Dr_vedl_doba_2_corr_hypoxie(ano) & Poloha_plodu_kod_vys(zahlavim) *** ANOTACE_zpusob_porodu(SC)
15	1	0.933	Bmi_nyni_int(nadvaha) & Bmi_pred_int(norma) & Dr_vedl_doba_2_corr_hypoxie(ano) *** ANOTACE_zpusob_porodu(SC)
16	5	0.933	Bmi_nyni_int(nadvaha) & Dr_vedl_doba_2_corr_hypoxie(ano) *** ANOTACE_zpusob_porodu(SC)
17	2	0.930	Bmi_nyni_int(nadvaha) & Bmi_pred_int(norma) & Dr_vedl_doba_2_corr_hypoxie(ano) & Poloha_plodu_kod_vys(zahlavim) *** ANOTACE_zpusob_porodu(SC)
18	6	0.928	Bmi_nyni_int(nadvaha) & Dr_vedl_doba_2_corr_hypoxie(ano) & Poloha_plodu_kod_vys(zahlavim) *** ANOTACE_zpusob_porodu(SC)
19	27	0.738	Bmi_nyni_int(obezita_1st) & Poloha_plodu_kod_vys(koncem_panevnm) *** ANOTACE_zpusob_porodu(SC)
20	21	0.734	Bmi_nyni_int(obezita_1st) & Bmi_pred_int(nadvaha) & Poloha_plodu_kod_vys(koncem_panevnm) *** ANOTACE_zpusob_porodu(SC)
21	50	0.734	Dr_vedl_doba_2_corr_diabetes(ano) *** ANOTACE_zpusob_porodu(SC)
22	41	0.718	Bmi_pred_int(nadvaha) & Poloha_plodu_kod_vys(koncem_panevnm) *** ANOTACE_zpusob_porodu(SC)
23	51	0.674	Dr_vedl_doba_2_corr_diabetes(ano) & Poloha_plodu_kod_vys(zahlavim) *** ANOTACE_zpusob_porodu(SC)
24	8	0.662	Bmi_nyni_int(nadvaha) & Poloha_plodu_kod_vys(koncem_panevnm) *** ANOTACE_zpusob_porodu(SC)
25	55	0.661	Poloha_plodu_kod_vys(koncem_panevnm) *** ANOTACE_zpusob_porodu(SC)
26	4	0.661	Bmi_nyni_int(nadvaha) & Bmi_pred_int(norma) & Poloha_plodu_kod_vys(koncem_panevnm) *** ANOTACE_zpusob_porodu(SC)
27	45	0.642	Bmi_pred_int(norma) & Poloha_plodu_kod_vys(koncem_panevnm) *** ANOTACE_zpusob_porodu(SC)
28	12	0.600	Bmi_nyni_int(norma) & Bmi_pred_int(norma) & Poloha_plodu_kod_vys(koncem_panevnm) *** ANOTACE_zpusob_porodu(SC)
29	16	0.599	Bmi_nyni_int(norma) & Poloha_plodu_kod_vys(koncem_panevnm) *** ANOTACE_zpusob_porodu(SC)

Figure 11: Results of Lisp-Miner task with Fisher quantifier, antecedent is selected attributes (not presented in diploma thesis) of length from 1 to 5 and succedent is class “SC” of method of childbirth.

Actual group of hypotheses: All hypothesis

Hypotheses in group: 6 Shown hypotheses: 6 Highlighted: 0

Nr.	Id	Conf	Hypothesis
1	6	0.928	Porody_celk(5) *** ANOTACE_zpusob_porodu(SpontVag)
2	5	0.887	Porody_celk(3) *** ANOTACE_zpusob_porodu(SpontVag)
3	4	0.870	Porody_celk(2) *** ANOTACE_zpusob_porodu(SpontVag)
4	3	0.859	Porody_celk(1) *** ANOTACE_zpusob_porodu(SpontVag)
5	2	0.215	Porody_celk(0) *** ANOTACE_zpusob_porodu(SC)
6	1	0.044	Porody_celk(0) *** ANOTACE_zpusob_porodu(OperVag)

Figure 12: Results of Lisp-Miner task with Fisher quantifier, antecedent is attribute “Porody_celk” and the succedent all classes of method of childbirth.

6.2.4 Association rules, Question 4

This task inspects an influence of all classes in dataset on class “vetsi4000” of annotation “ANOTACE_hmot_nov” (macrosomnia of fetus). Task setup is Fisher quantifier with base 100 instances, alpha level 0.001 and length of antecedent from 1 to 10. Succedent is class “vetsi4000” of the annotation.

Results are in Figure 13.

Actual group of hypotheses: All hypothesis

Hypotheses in group: 1310 Shown hypotheses: 1310 Highlighted: 0

Nr.	Id	Conf	Hypothesis
1	466	0.372	Bmi_nyni_int(obeziata_1st) & Hm_placenta_int(600_1000) & Hm_rozdil_int(16_25) & Porody_celk(1) *** ANOTACE_hmot_nov(vetsi4000)
2	464	0.371	Bmi_nyni_int(obeziata_1st) & Hm_placenta_int(600_1000) & Hm_rozdil_int(16_25) & Kod_pohl_nov_vys(chlapec) *** ANOTACE_hmot_nov(vetsi4000)
3	571	0.359	Bmi_nyni_int(obeziata_2st) & Hm_placenta_int(600_1000) & Kod_pohl_nov_vys(chlapec) *** ANOTACE_hmot_nov(vetsi4000)
4	1022	0.356	Hm_placenta_int(600_1000) & Hm_rozdil_int(16_25) & Kod_pohl_nov_vys(chlapec) & Porody_celk(1) *** ANOTACE_hmot_nov(vetsi4000)
5	415	0.356	Bmi_nyni_int(obeziata_1st) & Bmi_pred_int(norma) & Hm_placenta_int(600_1000) & Hm_rozdil_int(16_25) & Kod_pohl_nov_vys(chlapec) *** ANOTACE_hmot_nov(vetsi4000)
6	478	0.352	Bmi_nyni_int(obeziata_1st) & Hm_placenta_int(600_1000) & Kod_pohl_nov_vys(chlapec) & Porody_celk(1) & Tk_dia_int(optim) *** ANOTACE_hmot_nov(vetsi4000)
7	477	0.345	Bmi_nyni_int(obeziata_1st) & Hm_placenta_int(600_1000) & Kod_pohl_nov_vys(chlapec) & Porody_celk(1) *** ANOTACE_hmot_nov(vetsi4000)
8	1023	0.341	Hm_placenta_int(600_1000) & Hm_rozdil_int(16_25) & Kod_pohl_nov_vys(chlapec) & Porody_celk(1) & Tk_dia_int(optim) *** ANOTACE_hmot_nov(vetsi4000)
9	728	0.339	Bmi_pred_int(norma) & Hm_placenta_int(600_1000) & Hm_rozdil_int(16_25) & Kod_pohl_nov_vys(chlapec) & Porody_celk(1) *** ANOTACE_hmot_nov(vetsi4000)
10	368	0.333	Bmi_nyni_int(obeziata_1st) & Bmi_pred_int(nadvaha) & Hm_placenta_int(600_1000) & Kod_pohl_nov_vys(chlapec) & Porody_celk(1) *** ANOTACE_hmot_nov(vetsi4000)
11	361	0.331	Bmi_nyni_int(obeziata_1st) & Bmi_pred_int(nadvaha) & Hm_placenta_int(600_1000) & Hm_rozdil_int(16_25) *** ANOTACE_hmot_nov(vetsi4000)
12	482	0.330	Bmi_nyni_int(obeziata_1st) & Hm_placenta_int(600_1000) & Kod_pohl_nov_vys(chlapec) & Tk_syst_int(optim) *** ANOTACE_hmot_nov(vetsi4000)
13	1024	0.327	Hm_placenta_int(600_1000) & Hm_rozdil_int(16_25) & Kod_pohl_nov_vys(chlapec) & Tk_dia_int(norma) *** ANOTACE_hmot_nov(vetsi4000)
14	1035	0.327	Hm_placenta_int(600_1000) & Hm_rozdil_int(16_25) & Porody_celk(1) & Tk_syst_int(optim) *** ANOTACE_hmot_nov(vetsi4000)
15	467	0.325	Bmi_nyni_int(obeziata_1st) & Hm_placenta_int(600_1000) & Hm_rozdil_int(16_25) & Tk_dia_int(optim) *** ANOTACE_hmot_nov(vetsi4000)
16	597	0.325	Bmi_pred_int(nadvaha) & Hm_placenta_int(600_1000) & Hm_rozdil_int(16_25) *** ANOTACE_hmot_nov(vetsi4000)
17	416	0.324	Bmi_nyni_int(obeziata_1st) & Bmi_pred_int(norma) & Hm_placenta_int(600_1000) & Kod_pohl_nov_vys(chlapec) *** ANOTACE_hmot_nov(vetsi4000)
18	437	0.321	Bmi_nyni_int(obeziata_1st) & Dr_vedl_doba_1_corr_indukce(ano) & Hm_placenta_int(600_1000) & Kod_pohl_nov_vys(chlapec) *** ANOTACE_hmot_nov(vetsi4000)
19	436	0.318	Bmi_nyni_int(obeziata_1st) & Dr_vedl_doba_1_corr_indukce(ano) & Hm_placenta_int(600_1000) & Hm_rozdil_int(16_25) *** ANOTACE_hmot_nov(vetsi4000)
20	1026	0.318	Hm_placenta_int(600_1000) & Hm_rozdil_int(16_25) & Kod_pohl_nov_vys(chlapec) & Tk_dia_int(optim) & Tk_syst_int(optim) *** ANOTACE_hmot_nov(vetsi4000)
21	463	0.318	Bmi_nyni_int(obeziata_1st) & Hm_placenta_int(600_1000) & Hm_rozdil_int(16_25) *** ANOTACE_hmot_nov(vetsi4000)
22	607	0.318	Bmi_pred_int(nadvaha) & Hm_placenta_int(600_1000) & Kod_pohl_nov_vys(chlapec) & Porody_celk(1) *** ANOTACE_hmot_nov(vetsi4000)
23	1028	0.316	Hm_placenta_int(600_1000) & Hm_rozdil_int(16_25) & Kod_pohl_nov_vys(chlapec) & Tk_syst_int(optim) *** ANOTACE_hmot_nov(vetsi4000)
24	888	0.311	Dr_vedl_doba_1_corr_indukce(ano) & Hm_placenta_int(600_1000) & Hm_rozdil_int(16_25) & Kod_pohl_nov_vys(chlapec) *** ANOTACE_hmot_nov(vetsi4000)
25	483	0.311	Bmi_nyni_int(obeziata_1st) & Hm_placenta_int(600_1000) & Kod_pohl_nov_vys(chlapec) & Tk_syst_int(vysoky) *** ANOTACE_hmot_nov(vetsi4000)
26	479	0.310	Bmi_nyni_int(obeziata_1st) & Hm_placenta_int(600_1000) & Kod_pohl_nov_vys(chlapec) & Tk_dia_int(norma) *** ANOTACE_hmot_nov(vetsi4000)
27	367	0.309	Bmi_nyni_int(obeziata_1st) & Bmi_pred_int(nadvaha) & Hm_placenta_int(600_1000) & Kod_pohl_nov_vys(chlapec) *** ANOTACE_hmot_nov(vetsi4000)
28	475	0.309	Bmi_nyni_int(obeziata_1st) & Hm_placenta_int(600_1000) & Kod_pohl_nov_vys(chlapec) *** ANOTACE_hmot_nov(vetsi4000)
29	414	0.309	Bmi_nyni_int(obeziata_1st) & Bmi_pred_int(norma) & Hm_placenta_int(600_1000) & Hm_rozdil_int(16_25) *** ANOTACE_hmot_nov(vetsi4000)
30	875	0.307	Bmi_pred_int(obeziata_1st) & Hm_placenta_int(600_1000) & Kod_pohl_nov_vys(chlapec) *** ANOTACE_hmot_nov(vetsi4000)
31	1032	0.305	Hm_placenta_int(600_1000) & Hm_rozdil_int(16_25) & Porody_celk(1) *** ANOTACE_hmot_nov(vetsi4000)

Figure 13: Results of Lisp-Miner task with Fisher quantifier, antecedent is selected attributes (not presented in diploma thesis) of length from 1 to 10 and succedent is class “vetsi4000” of weight of fetus.

6.3 Discussion

I explain the principle howto interpret results from Chapter 6.2 on Question 1, Figure 8 and Figure 9. Then I interpret results for others questions (Question 2, Question 3, Question 4) .

6.3.1 Question 1

Figure 8 shows interesting features (blue mark), which indicate influence on positive outcome of delivery (stated via good apgar sum in 5th minute). I would expect that at least finished 36th pregnancy week (attribute “tt_dokonceny36_int(vetsi36)”) is good for newborn. I do not know anything about influence of previous delivery. Very important is knowledge that obstetrics medicaments (oxytocin and spasmolytics - – attributes “dr_vedl_doba_2_corr_oxytocin“, “dr_vedl_doba_1_corr_spasmolytika“) and umbilical cord

around neck of fetus (attribute “pozn_corr_pupecnikkk“) do not damage positive outcome. What could be surprising is that streptococcus also not damage positive outcome.

All of these hypotheses are based only on one metric – confidence in the association rule (“Conf“ heading in Figure 8 or Figure 9). It is not enough to properly describe the association rule, thus I use next metric called coverage. Chapter 2.5.5 explains terms confidence and coverage. Figure 9 serves to justify formulated hypotheses and to put them into context.

To provide an example, I choose hypothesis of streptococcus.

Simple association rule is “Pozn_corr_sag(ano) => ANOTACE_apgar_suma_5(vetsi7)“ and its 4-fold table including different Lisp-Miner metrics is shown in Figure 14.

Confidence is 99.2% and it is stated in Lisp-Miner also as confidence or validity of association rule. What I state in my diploma thesis as coverage, it is “Completness“ in Lisp-Miner (“Cmplt“ sign). Coverage of this rule is 16.1%.

This association rule is very probable valid, because it has high confidence and covers approximately 1/6 of all possible situations of the class “vetsi7“ of the annotation. Another significant characteristic of this rule is the number of instances in “a“ and “b“ values. “a“ value represents 5 818 instances of rule validity in comparison to “b“ value (no validity of rule) of 46 instances, which is big difference.

My conclusion is that I would trust to this association rule. What's more, it confirms the hypothesis of no harm of streptococcus on positive outcome of apgar sum in 5th minute (no harm on positive outcome of delivery).

We see from Figure 9 the space to find another association rule, which may be even better and more valuable. I examine association rules of Figure 15, Figure 16 and Figure 17. 4-fold tables of Figure 15, Figure 16 and Figure 17 are snapshots of results exported to web page by Lisp-Miner module “4ftResult.exe“.

Hypothesis					
Antecedent: Pozn_corr_sag(ano)					
Succedent: ANOTACE_apgar_suma_5(vetsi7)					
Condition: (empty)					
TEXT	DATA	MAP/GRAPH	PIE	BAYES	AR2NL
Contingency table					
	Succedent	NOT Succedent			
Antecedent	5818	46	5864		
NOT Antecedent	30343	565	30908		
	36161	611	36772		
Values from contingency table:					
a	5818	5818	a-frequency from the contingency table		
b	46	46	b-frequency from the contingency table		
c	30343	30343	c-frequency from the contingency table		
d	565	565	d-frequency from the contingency table		
r	5864	5864	r-frequency (a+b) from the contingency table		
n	36772	36772	n-frequency (a+b+c+d) from the contingency table		
Conf	0.99	0.9921555252	Confidence (validity): $a/(a+b)$		
DConf	0.16	0.1606871599	D-Confidence: $a/(a+b+c)$		
EConf	0.17	0.1735831611	E-Confidence: $(a+d)/(a+b+c+d)$		
Supp	0.16	0.1582182095	Support: $a/(a+b+c+d)$		
Cmplt	0.16	0.1608915683	Completeness: $a/(a+c)$		
AvgDf	0.01	0.0089196365	Average difference: $a(a+b+c+d)/((a+b)(a+c))-1$		
LBound	0	0	Lower bound implication ($p=0.9$)		
UBound	1	1	Upper bound implication ($p=0.9$)		
ELBound	1	0.9999999996	Lower bound equivalence ($p=0.9$)		
EUBound	0	0	Upper bound equivalence ($p=0.9$)		
DLBound	1	0.9999999995	Lower bound double implication ($p=0.9$)		
DUBound	0	0	Upper bound double implication ($p=0.9$)		
Fisher	0	0	Fisher test		
Chi-Sq	32.85	32.8498839424	Chi-square test		
bMean	0.99	0.9919877259	Bayesian Mean		

Figure 14: Simple association rule of streptococcus with 4-fold table and some of all possible Lisp-Miner metrics.

	Succedent	¬Succedent
Antecedent	1283	7
¬Antecedent	39962	622

Figure 15: 4-fold table of association rule of antecedent streptococcus and oxytocin. Confidence is 99.5% and coverage 3.1%.

	Succedent	¬Succedent
Antecedent	1247	4
¬Antecedent	40015	625

Figure 16: 4-fold table of association rule of antecedent streptococcus, oxytocin and pregnancy week. Confidence is 99.7% and coverage 3.0%.

	Succedent	–Succedent
Antecedent	460	0
–Antecedent	41352	631

Figure 17: 4-fold table of association rule of antecedent streptococcus, oxytocin, pregnancy week and number of previous childbirths. Confidence is 100% and coverage 1.1%.

From these results, it is obvious that additional class of attribute improves the model, but reduce the coverage. Oxytocin, pregnancy week or number of previous deliveries work as adjustment in different situations. Other results confirm that spasmolytics help to improve confidence as well.

It is very difficult to say, which association is better due to diverse coverage. I would use them individually according to needs of obstetricians. I can compare for example association rules of Figure 15 and Figure 16 because they cover similar instances. Here, I would recommend association rule of Figure 16 because of slightly better confidence.

The best model according to confidence is association rule of Figure 17:

- `Dr_vedl_doba_2_corr_oxytocin(ano) & Porody_celk(1) & Pozn_corr_sag(ano) & Tt_dokonceny36_int(vetsi36) => ANOTACE_apgar_suma_5(vetsi7).`

The interpretation of such rule is that the apgar sum in 5th minute is all right (then doctor expects no complication in childbirth) if a doctor applies oxytocin within 2nd stage of labour to support contractions of gravida, who has to be at least in 36th pregnancy week and already having 1 child. Previous streptococcus of gravida does not deteriorate apgar sum.

I could even say that having streptococcus in this combination better confidence of positive outcome of 460 gravidas from set of 3 620 gravidas recorded in association rule of “Id 11” of Figure 9. The reason underlying such hypothesis should be investigated.

The second result of this section is Figure 10. It concentrates on negative outcome via annotation of pH measure of newborn.

Results are not as convincing as in former case, but I think they are also interesting. The main impact on the negative outcome (pH lower 7.15) has deceleration of fetus registered in 2nd stage of labour (attribute “dr_vedl_doba_2_corr_decelerace”). Combinations of other classes improve confidence in association rule with deceleration of fetus.

I depict one association rule from Figure 10 with confidence 55.4% and coverage 5%, “a” value 129 instances and “b” value 104 instances:

- Dr_vedl_doba_1_corr_epidural(ano) & Dr_vedl_doba_1_corr_indukce(ano)
& Dr_vedl_doba_2_corr_decelerace(ano) & Porody_celk(0)
=> ANOTACE_ph_7_15(mensi7_15).

Interpretation of this rule is that pH of newborn is bad (then doctor expects complications in childbirth) if a doctor applies epidural and induction on gravida within 1st stage of labour, who never had a child, and registers deceleration of fetus in 2nd stage of labour.

In other words, it means that the doctor wants to relieve gravida (having no child till now) from pain and to support finishing of delivery. Afterwards, when there is decrease of heart beat of fetus, doctors should expect low pH of fetus with 55.4% chance.

6.3.2 Question 2

The biggest influence on Caesarean method of delivery has hypoxia of fetus (attribute “dr_vedl_doba_2_corr_hypoxie”) with confidence 94.7% and coverage 13.5%. Present diabetes of gravida within 2nd stage of labour (attribute “dr_vedl_doba_2_corr_diabetes”) has confidence 73.4% and coverage 2.3%. Position of fetus (breech position, attribute “poloha_plodu_kod_vys”) has confidence 66.1% and coverage 20%.

Association rule of hypoxia of fetus is:

- Dr_vedl_doba_2_corr_hypoxie(ano) => ANOTACE_zpusob_porodu(SC).

The interpretation is easy, there is 94.7% chance (991 instances) that delivery ends via Caesarean section, when doctor registers hypoxia of fetus within 2nd stage of labour (fetus does not have enough oxygen).

Combinations of classes (see results in Figure 11) show improved confidence of association rules with hypoxia as antecedent and with Caesarean section as succedent, when current BMI of gravida is present in antecedent. Thus I present association rule:

- Bmi_nyni_int(obezita_1st) & Dr_vedl_doba_2_corr_hypoxie(ano)
=> ANOTACE_zpusob_porodu(SC).

This association rule has confidence 96.9%, coverage 2.9%, “a” value 217 instances and “b” value 7 instances.

The interpretation is that a child is going to be delivered by Caesarean section if gravida has obesity of 1st degree (measured by BMI before delivery) and doctor registers hypoxia of fetus within 2nd stage of labour.

Instances of Caesarean section are equally distributed among classes, therefore most of association rules of Figure 11 have rather descriptive character.

6.3.3 Question 3

Figure 12 shows results of number of previous childbirths before the current childbirth and their impact on method of current delivery. It seems that having more children before current delivery has positive influence. The other way round, nulliparous women are liable to Caesarean or operational delivery.

Due to small confidence, I rather falsify hypotheses of nulliparous women.

Regarding women with children, it is fact that having more children lead to positive outcome of delivery (spontaneous delivery).

I mention 2 association rules with high confidence and coverage:

- $\text{Porody_celk}(2) \Rightarrow \text{ANOTACE_zpusob_porodu}(\text{SpontVag}),$

(Confidence 87%, coverage 10%, “a” value 3 457 instances, “b” value 518 instances),

- $\text{Porody_celk}(1) \Rightarrow \text{ANOTACE_zpusob_porodu}(\text{SpontVag}),$

(Confidence 85.9%, coverage 40.8%, “a” value 14 090 instances, “b” value 2 306 instances).

I trust most to the second association rule because it has high coverage, high “a” value and almost the same confidence as first association rule. Higher confidence is prove that having more children leads to spontaneous childbirth. However, there can be also any not known process underlying this explanation.

The interpretation of the second association rule is that gravida is going to deliver child spontaneously with 85.9% chance, if she already had one child.

6.3.4 Question 4

Experiments with succedent macrosomia of fetus prove that the biggest impact have current BMI of gravida (attribute “bmi_nyni_int“, classes obesity of 1st and 2nd degree), BMI of gravida before pregnancy (attribute “bmi_pred_int“, classes overweight and obesity of 1st degree) and weight of placenta (attribute “hm_placenta_int“, category of 600 to 1 000 grams).

Combination of attributes (see results in Figure 13) slightly improves confidence, e.g. BMI of gravida before the childbirth (class obesity of 1st degree) improves confidence from 16% as single attribute to 37% in case of hypothesis in Figure 13 with “Id 466“. It is not possible to mention any of these association rules, which could significantly influence macrosomia of fetus, due to very low confidence.

7 Evaluation of results via other data mining method

I evaluate attributes of association rules from Chapter 6 by Random Forest classifier implemented in Orange software.

Random Forest classifier produces set of classification trees based on the training data. Every tree arbitrarily chooses attributes and use the best one for the split of the individual classification tree. The resulting class is the most frequent one from all developed classification trees.[19]

I choose Random Forest because the source [11] (see Chapter 4.2) states that it performs very well on highly imbalanced data, which is exactly the case of CTU Database.

To prove that, I make the Nearest neighbour and Random Forest classifications of attributes (dr_vedl_doba_2_corr_oxytocin, porody_celk, pozni_corr_sag and tt_dokonceny36_int) with 10-fold cross-validation and classification target ANOTACE_apgar_suma_5(vetsi7). This combination of attributes and one annotation comes out of the association rule:

■ Dr_vedl_doba_2_corr_oxytocin(ano) & Porody_celk(1) & Pozni_corr_sag(ano) & Tt_dokonceny36_int(vetsi36) => ANOTACE_apgar_suma_5(vetsi7).

The process of Orange software is in Figure 18, default setup of classifiers shows Figure 19. Accuracy, sensitivity and specificity of classifications are in Table 5 or in Appendix 4.

We see from Table 5 that Random Forest with 10 individually developed trees performs slightly better than 5-Nearest neighbour classifier. More important is the fact that Random Forest classification is much faster than Nearest neighbour classifier.

Table 6 shows accuracy, sensitivity and specificity of Random Forest classification for attributes and annotations according to association rules presented in Chapter 6.

Lines marked in grey of Table 5 and Table 6 represent classification of entire dataset. White lines are classifications of selected attributes.

We see that selected attributes classify almost the same (slightly lower classification accuracy) as entire dataset except annotation of method of delivery. Attribute of current BMI of gravida (bmi_nyni_int) has no significance for this annotation.

When I classify attributes to positive outcome (normal states), selected attributes increase or almost equal the sensitivity of entire dataset. Classification to negative outcome (pathological states) increases or equals specificity.

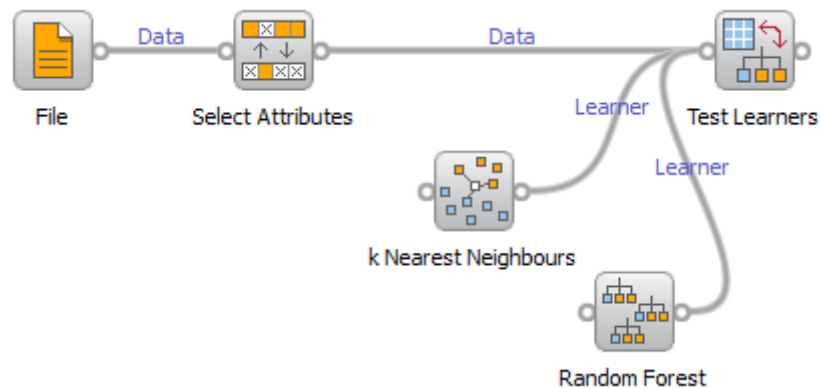


Figure 18: Classification process of Orange software. The process opens dataset with proper annotation, selects attributes according to association rules and do the classification of Random Forest and Nearest neighbour classifiers.

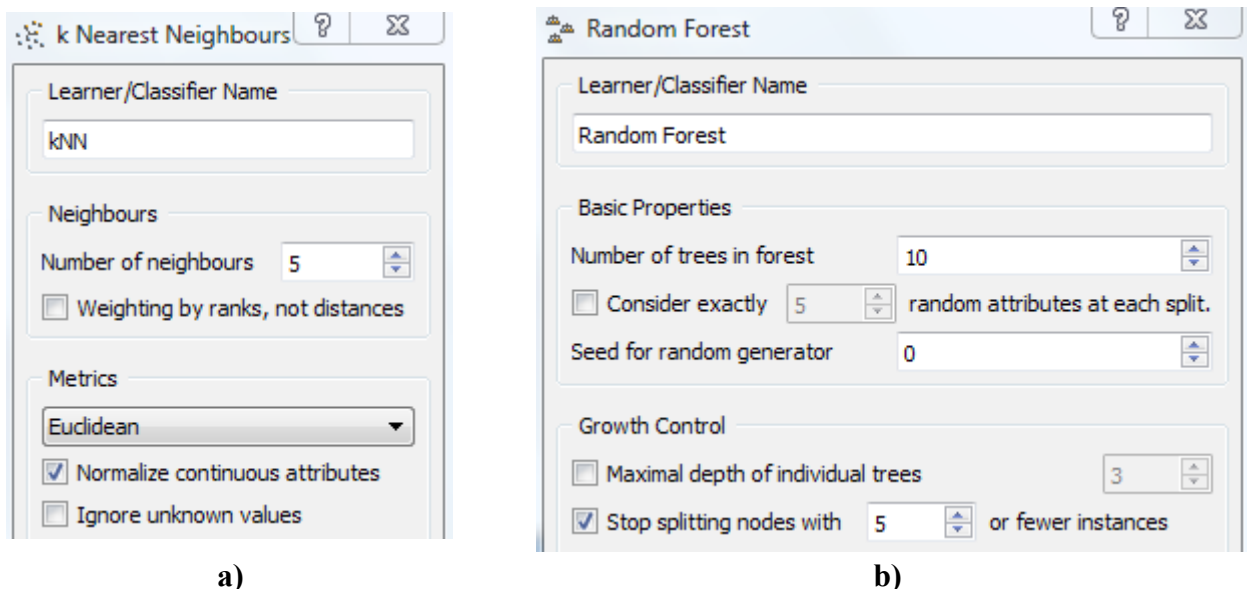


Figure 19: Orange software - default setup of a) 5-Nearest neighbour classifier and b) 10 individually developed trees in Random Forest classifier.

Table 5: Classification of 5-nearest neighbour classifier and Random Forest classifier with 10 individually developed trees. Lines Marked in grey are classifications of entire dataset and white lines represent classifications of selected attributes.

Classifier (Orange Software)	Classification accuracy [%]	Sensitivity [%]	Specificity [%]
k Nearest Neighbours (k = 5)	98.52	99.89	6.64
	98.57	99.88	11.06
Random Forest (10 trees)	98.62	99.99	6.64
	98.58	99.98	4.58

Table 6: Classifications of Random Forest classifier with 10 individually developed trees. Selected attributes and annotations represent classifications of relevant attributes of association rules from Chapter 6.

Question	Selected attributes	Classification target Annotation(class)	Classification accuracy [%]	Sensitivity [%]	Specificity [%]
1	Entire dataset	ANOTACE_apgar_suma_5(vetsi7)	98.62	99.99	6.64
	pozn_corr_sag	ANOTACE_apgar_suma_5(vetsi7)	98.53	100.00	0.00
	dr_vedl_doba_2_corr_oxytocin, porody_celk, pozn_corr_sag, tt_dokonceny36_int	ANOTACE_apgar_suma_5(vetsi7)	98.58	99.98	4.58
	Entire dataset	ANOTACE_ph_7_15(mensi7_15)	81.62	3.31	99.37
	dr_vedl_doba_1_corr_epidural, dr_vedl_doba_1_corr_indukce, dr_vedl_doba_2_corr_decelerace, porody_celk	ANOTACE_ph_7_15(mensi7_15)	81.52	0.00	100.00
2	Entire dataset	ANOTACE_zpusob_porodu(SC)	93.83	80.57	99.56
	dr_vedl_doba_2_corr_hypoxie	ANOTACE_zpusob_porodu(SC)	82.50	13.36	99.85
	bmi_nyni_int, dr_vedl_doba_2_corr_hypoxie	ANOTACE_zpusob_porodu(SC)	82.50	13.36	99.85
3	Entire dataset	ANOTACE_zpusob_porodu(SpontVag)	93.83	99.68	70.59
	porody_celk	ANOTACE_zpusob_porodu(SpontVag)	80.25	100.00	0.00
4	Entire dataset	ANOTACE_hmot_nov(vetsi4000)	85.69	0.13	99.98
	bmi_nyni_int, bmi_pred_int, hm_placenta_int	ANOTACE_hmot_nov(vetsi4000)	84.46	0.53	99.97

8 Encountered problems and possible future solution

I identified two technical problems. One relates to creating data and the second one to actual DM, to be precise to computing power. In the end of this chapter, I add one non-technical problem. In my opinion, it is also important here to point out the problem with medical data interpretation and associations.

Prior to dataset creation, it is necessary to extract proper attributes and proper instances from a database. The extraction of attributes from CTU database is done by SQL scripts. I organized SQL scripts into files, which is more comfortable while handling that in a software, e.g. RapidMiner. In such a large database, it is very problematic to write these files and follow conditions of attributes or naming of attributes and tables. I have no solution for this problem, because it is given by structure of PostgreSQL database management system.

Lisp-Miner needs dataset in a form of “.mdb” file (file of Microsoft Access database management system, known as Microsoft Access). **Creating dataset** from RapidMiner to the Microsoft Access database is easy, if we overcome first painful contact with these softwares. RapidMiner cannot create straightforward new Microsoft Access database. To solve this I need to create empty database and point to the database from RapidMiner. Versions of RapidMiner and MS Access has to comply, it means if we use 32-bit version of Microsoft Access, we need to install 32-bit version of RapidMiner. Another problem is that creating database file of the dataset is very time-consuming. All of these problems may solve evolution of Lisp-Miner called “Ferda Data Miner”. This software should directly connect CTU database and Lisp-Miner methods and settings.

Computing power is crucial parameter in data mining. Before running a task to obtain results I had to face the problem how many attributes to choose and howto setup tasks. Every manual intervention may lead to improper results, more attributes lead to more time spend on the task and thus we would inspect lower amount of possible setup of tasks.

To balance time consumption in DM software mentioned in Chapter 3 I would recommend to have less than 20 attributes, rather about 10 most significant attributes for target class. To do first feature extraction of large datasets as CTU Database is, I would recommend to have at least multicore computers or to have computer grid.

I give one example, I run task in Lisp-Miner, which lasted approximately 14 hours and a half on my ordinary laptop with 2-core processor. When I have the chance to test the same task on available Biodat Research Group 8-core processor computer, it lasted 8 hours and a half.

Other setup, e.g. extraction of interesting relations from set of around 60 attributes in RadViz of Orange software may last from many days to weeks till finish (dependent on settings). Despite the fact that the task is possible to stop and get some of results, to inspect all possible combinations is not feasible even with multicore processor computers. Here I would use computer grid.

The **non-technical problem** relates to medicine domain. A technician, who have not touched obstetrics field or without medical background, has many problems to find valuable attributes or association rules and subsequently to interpret them.

It could be valuable to have obstetrician next to a technician, who helps with these topics, especially with association rules results.

In my case, within my work we determine roughly 100 interesting attributes. Then we get feedback from obstetricians of hospital of FN Brno, which attributes they consider significant. I used theirs selection in next DM. Nevertheless, DM specialist should be careful and should check attributes once again to avoid mistakes.

When I see results of association rules, I do not consider most of association rules as interesting, because I miss the insight into the obstetric problematic. My supervisor helped me with the selection, because I did not have enough time (because of diploma thesis deadline) to provide results to obstetrician.

**CONCLUSION,
BIBLIOGRAPHY,
APPENDICES**

9 Conclusion

DM is the identification of interesting information or knowledge, which is hidden in data. DM techniques are enough reliable, understandable and highly accurate. That is why they are widely used in many fields.

Theoretical part concentrates on general overview of DM including two interesting visualization techniques (Nomogram and RadViz), software tools and research in medical domain.

Software tools are used to help me familiarize with the CTU database (SchemaSpy) and to perform data preparation (RapidMiner), association rules (Lisp-Miner) and evaluation by classification (Orange software). I also describe Weka software. I do not use it, because previously mentioned softwares handle all of related problems.

I made a research of interesting features in DM of pregnancies and deliveries with the help of SpringerLink search engine. I described similar databases as is CTU Database (see Chapter 4).

Preeclampsia is a cause of preterm birth, infection or inflammation influence preterm birth, bacterial vaginosis increase risk of low weight of fetus, amniotic infection and preterm birth. BMI of gravida should be somehow associated with birth weight of fetus. Also higher BMI increase risk of preeclampsia, hypertension and complicated birth (as Caesarean section or induction) (see Chapter 4 or sources [6, 7]).

Interesting datasets related to childbirth are Aberdeen Maternity and Neonatal Databank, Missouri maternal dataset and Danish National Birth Cohort. They use chi-square statistical test or BestFirst search algorithm to extract features. They evaluate risk by odd ratio metric or by regression analysis. The biggest found dataset is United States Nationwide Inpatient Sample of almost 8 million hospital records (not related to pregnancy). Here, they promote classification method Random Forest with random sub-sampling to effectively handle large datasets of imbalanced classes (see Chapter 4 or sources [8, 9, 10, 11]).

Practical part starts with description of CTU Database, which is based on PostgreSQL database management system, access via “brnodata_reader“. Chapter 5 also describes selected annotations and features for DM, which are then used to interpret association rules in Chapter 6 – Experiments.

Chapter 6 – Experiments is the actual part of data preparation and association rules. Data preparation was based on manual selection of attributes, e.g. removal of redundant attributes (multiple attributes of the same basis, extremely imbalanced or empty classes). Limitations shrink the dataset of 4000 instances, because I want to avoid errors of data. New attributes (BMI) are created and most of attributes are categorized to intervals due to association rules of Lisp-Miner software.

I created datasets in RapidMiner software. Attribute “ic_chor” is necessary for Lisp-Miner as unique identifier of medical records.

I focused on 4 areas to obtain results of association rules – normal or pathological outcome of newborn (Question 1), factors that influence Caesarean section (Question 2), influence of number of previous deliveries on the current delivery (Question 3) and macrosomia of fetus (weight higher than 4000 grams; Question 4).

I made a selection of valuable association rules in Lisp-Miner in two steps. I chose Fisher quantifier, base value 100 (minimum “a” value of 4-fold table) and set antecedent of association rule to be length of 1 to 1 (maximum 1 attribute as antecedent), which provides interesting classes (from entire dataset) to appropriate class of annotation. The second step was to make combination of that best classes.

Thus I obtained many combinations of association rules (Figure 9 to Figure 13) with the confidence in association rule.

The results are that streptococcus, spasmolytics or oxytocin do not harm to normal outcome of newborn. Produced rules have mostly high confidence and low coverage. The very good association rule is “Pozn_corr_sag(ano) => ANOTACE_apgar_suma_5(vetsi7)“, with confidence 99.2% and coverage 16.1%. Deceleration of fetus influences most the negative outcome of newborn (annotation of pH).

Caesarean section is influenced most by hypoxia of fetus (confidence 94.7%, coverage 13.5%).

Association rules method proves that there is association between number of previous and ongoing spontaneous delivery. The more previous deliveries, the better confidence (and significantly lower coverage due to number of instances in CTU Database).

I identified that the BMI (obesity) and weight of placenta (600 to 1000 g) has the main impact on macrosomia of fetus (as mentioned in Chapter 4). However, the confidence

is too low to make any reasonable conclusion from these attributes.

I evaluated attributes of association rules by Random Forest classification method with 10 individually developed classification trees. I proved on one association rule that Random Forest performs slightly better than 5-Nearest neighbour classifier.

Classification of selected attributes has almost the same classification accuracy as classification of entire dataset. The sensitivity is equal or higher in classification to positive outcome. The specificity is equal or higher in classification to negative outcome.

Results of classification are understandable. Association rules explain very specific area that we want to discover. Classification focuses on the general prediction of entire problem. If we apply attributes identified by association rules to classification, we enlarge the space of instances (especially in imbalanced data) and therefore the space of possible misclassifications.

In the last chapter of this diploma thesis I identified the biggest problems – creating dataset into Microsoft Access database, low computing power and interpretation of association rules of obstetrics data. Possible solutions are to use “Ferda data miner“ instead of Lisp-Miner, use computer grids or multi-core processor computers instead of common laptop and be in contact with obstetricians as possible while explaining obstetrics association rules.

Bibliography

Notes:

¹ Electronic article (SpringerLink search engine)

² This website is a compilation of publications or electronic articles

³ Official website to download the software tool (software description is in Chapter 3)

- [1] MAŘÍK, Vladimír. Umělá inteligence. 1. vyd. Praha: Academia, 2003, 475 s. ISBN 80-200-1044-0.
- [2] WITTEN, I. Data mining: practical machine learning tools and techniques. 2nd ed. San Francisco: Morgan Kaufmann Publishers, 2005, 525 s. ISBN 01-208-8407-0.
- [3] KLÉMA, Jiří. Symbolické metody strojového učení – PRAVIDLA. 2006.
- [4] AZEVEDO, Paolo J. Comparing Rule Measures for Predictive Association Rules. [online]. 2007 [cited 2012 April 10]. Available from: <http://www3.di.uminho.pt/~pja/ps/conviction.pdf>
- [5] NOVÁKOVÁ, Lenka. Visualization data for Data Mining. Prague, 2009. Doctoral Thesis. Czech Technical University in Prague. Vedoucí práce Prof. Olga Štěpánková.
- [6]¹ Gravett MG, et al.: Global report on preterm birth and stillbirth (2 of 7): discovery science. BMC Pregnancy and Childbirth 2010, 10(Suppl 1):S2.

- [7]¹ FREDERICK, Ihunnaya O., Michelle A. WILLIAMS, Anne E. SALES, Diane P. MARTIN and Marcia KILLIEN. Pre-pregnancy Body Mass Index, Gestational Weight Gain, and Other Maternal Characteristics in Relation to Infant Birth Weight. *Maternal and Child Health Journal* [online]. 2008, Volume: 12, Issue: 5, Pages: 557-567 [cited 2012 April 30]. ISSN 1092-7875. DOI: 10.1007/s10995-007-0276-2. Available from: <http://www.springerlink.com/index/10.1007/s10995-007-0276-2>
- [8]¹ BHATTACHARYA, Sohinee, Doris M CAMPBELL, William A LISTON and Siladitya BHATTACHARYA. Effect of Body Mass Index on pregnancy outcomes in nulliparous women delivering singleton babies. *BMC Public Health* [online]. Volume: 7, Issue: 1, Pages: 168- [cited 2012 April 30]. ISSN 14712458. DOI: 10.1186/1471-2458-7-168. Available from: <http://www.biomedcentral.com/1471-2458/7/168>
- [9]¹ WHITEMAN, Valerie E., Muktar H. ALIYU, Euna M. AUGUST, Cheri MCINTOSH, Jingyi DUAN, Amina P. ALIO and Hamisu M. SALIHU. Changes in prepregnancy body mass index between pregnancies and risk of gestational and type 2 diabetes. *Archives of Gynecology and Obstetrics* [online]. 2011, Volume: 284, Issue: 1, Pages: 235-240 [cited 2012 April 30]. ISSN 0932-0067. DOI: 10.1007/s00404-011-1917-7. Available from: <http://www.springerlink.com/index/10.1007/s00404-011-1917-7>
- [10]¹ OROZOVA-BEKKEVOLD, Ivanka, Henrik JENSEN, Lone STENSBALLE and Jørn OLSEN. Maternal vaccination and preterm birth: using data mining as a screening tool. *Pharmacy World* [online]. 2007-5-23, Volume: 29, Issue: 3, Pages: 205-212 [cited 2012 April 30]. ISSN 0928-1231. DOI: 10.1007/s11096-006-9077-8. Available from: <http://www.springerlink.com/index/10.1007/s11096-006-9077-8>
- [11]¹ Khalilia et al.: Predicting disease risks from highly imbalanced data using random forest. *BMC Medical Informatics and Decision Making* 2011 11:51.
- [12]² Data Mining and Analytic Technologies (Kurt Thearling) [online]. 2010 [cited 2012 March 20]. Available from: <http://thearling.com/>
- [13]² Data mining. Wikipedia, the free encyclopedia [online]. 2012 [cited 2012 April 10]. Available from: http://en.wikipedia.org/wiki/Data_mining
- [14]² Cluster analysis. Wikipedia, the free encyclopedia [online]. 2012 [cited 2012 April 10]. Available from: http://en.wikipedia.org/wiki/Cluster_analysis

- [15]² Association rule learning. Wikipedia, the free encyclopedia [online]. 2012 [cited 2012 April 10]. Available from: http://en.wikipedia.org/wiki/Association_rule_learning
- [16]³ SchemaSpy [online]. 2010 [cited 2012 April 20]. Available from: <http://schemaspy.sourceforge.net/>
- [17]³ Rapid - I [online]. 2012 [cited 2012 April 20]. Available from: <http://rapid-i.com/>
- [18]³ The official site of the LISp-Miner project [online]. 2012 [cited. 2012 April 20]. Available from: <http://lispminer.vse.cz/>
- [19]³ Orange – Data Mining Fruitful & Fun [online]. 2012 [cited 2012 April 20]. Available from: <http://orange.biolab.si/>
- [20]³ Weka 3 - Data Mining with Open Source Machine Learning Software in Java [online]. 2012 [cited 2012 April 20]. Available from: <http://www.cs.waikato.ac.nz/ml/weka/>
- [21] Relační vs. objektově-relační vs. objektové databáze. Fakulta informatiky Masarykovy univerzity [online]. Unknown [cited 2012 February 20]. Available from: <http://www.fi.muni.cz/~xbatko/oracle/compare.html>
- [22] Výpočet BMI, Body Mass Index. Výpočet.cz [online]. 2007 [cited 2011 August 05]. Available from: <http://www.vypocet.cz/bmi>
- [23] Krevní tlak. Prevence nemoci a podpora zdraví [online]. 2004 [cited 2011 August 05]. Available from: <http://www.cba.muni.cz/prevencenemoci/modules.php?name=Content&pa=showpage&pid=10>
- [24] Datový standard MZČR [online]. 2008 [cited 2012 February 20]. Available from: <http://ciselniky.dasta.mzcr.cz/>

List of appendices

Appendix 1	...	Example of a simple decision tree
Appendix 2	...	Example of relations among tables of CTU Database generated by SchemaSpy software
Appendix 3	...	Example of possible data preparation and data transformation in RapidMiner
Appendix 4	...	Example of evaluation of the result from Chapter 6 in Orange software

Appendix 1

The pseudo example of a decision tree is shown in Figure 20.

For example we want to figure out the spending of a lot of money with classes “yes”, “no”. We have attributes salary and purchase.

The tree starts with the node “salary” with branches “less or equal 20 000”, “more than 20 000” (Czech crown per month). The successive node is “purchase” with branches “mobile phone”, “car”. In case of branch “less or equal 20 000” the prediction of spending a lot is “no”. In case of branch “car” prediction is “yes” and for branch “mobile phone” is “no”.

Interpretation of this decision tree is that people with higher salary spend a lot of money, when they buy a car.

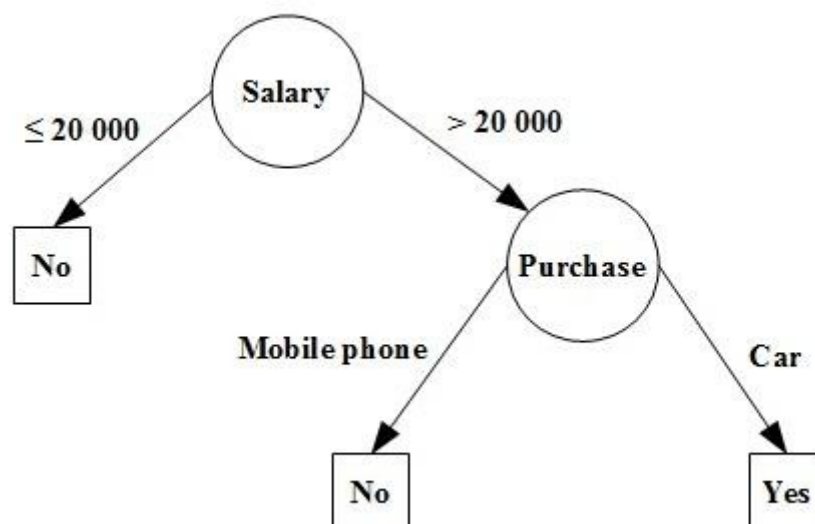


Figure 20: Example of a simple pseudo decision tree. Circles are nodes and predictions are in squares. Classes are above branches of the decision tree.

Appendix 2

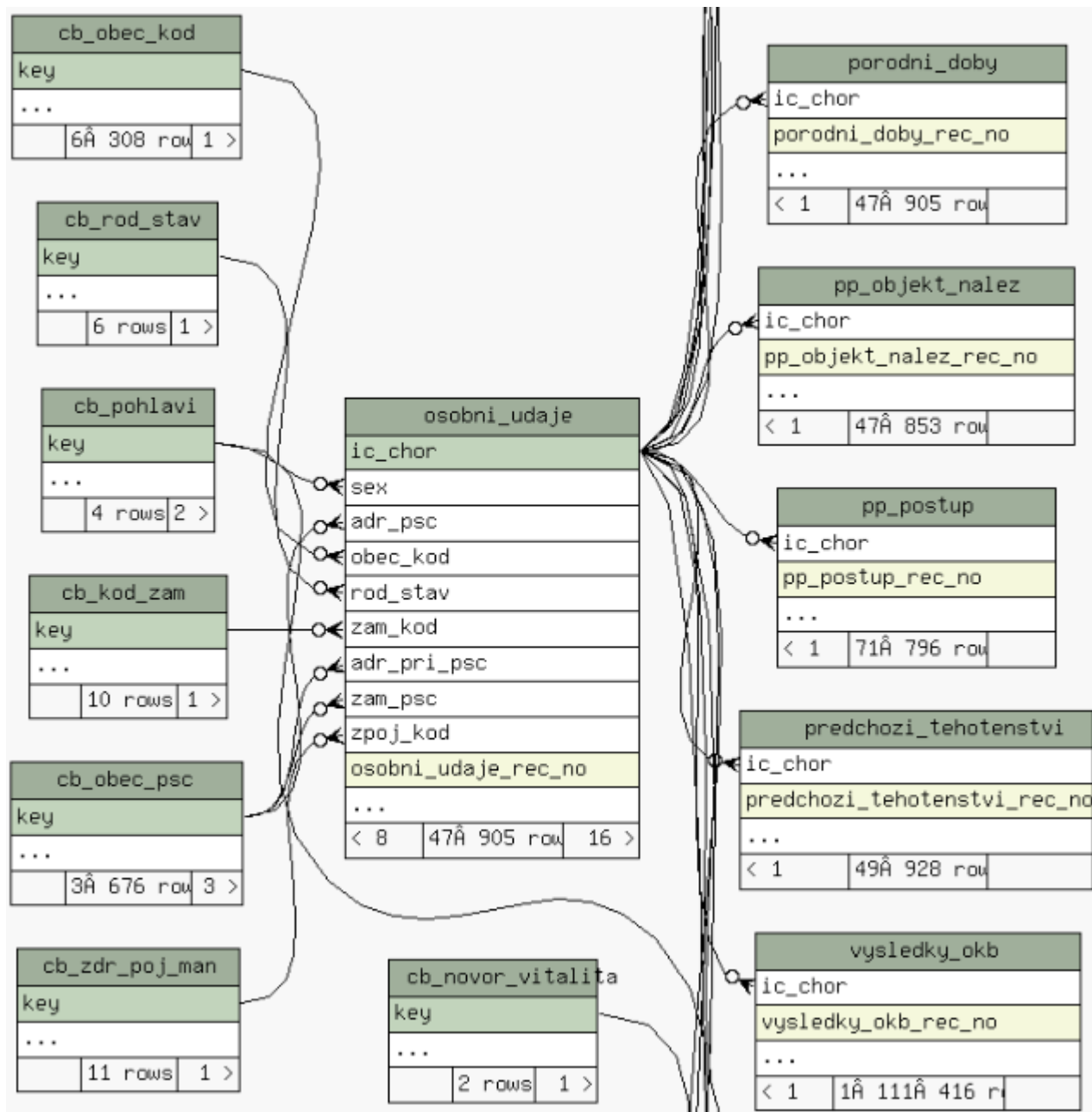


Figure 21: Example of relations among tables of CTU Database generated by SchemaSpy software. Understandably, the table personal data “osobni_udaje” has the most relations, because it contains specific attributes used in the rest of tables.

Appendix 3

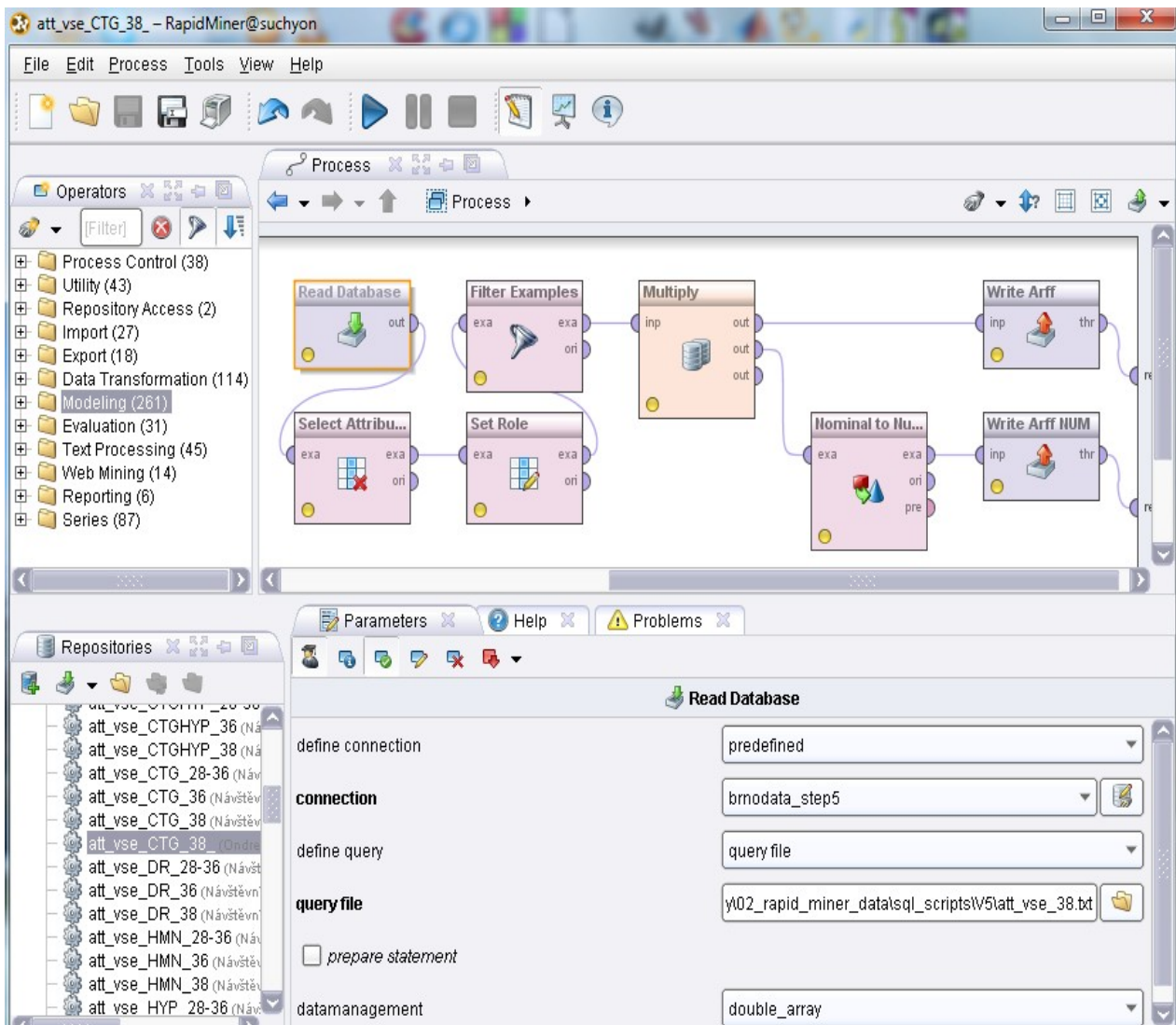


Figure 22: Example of possible data preparation and data transformation in RapidMiner 5.2. The process read from CTU database according to settings in “Parameters” window. Further settings of the database and other operators is not shown. Then it selects attributes for annotation “ANOTACE_ctg-spatny” and set it as a label (special attribute) from regular one. It filters missing values of the annotation and write Weka “.arff” file in nominal and in numerical form (“Nominal to Numerical” operator is used).

Appendix 4

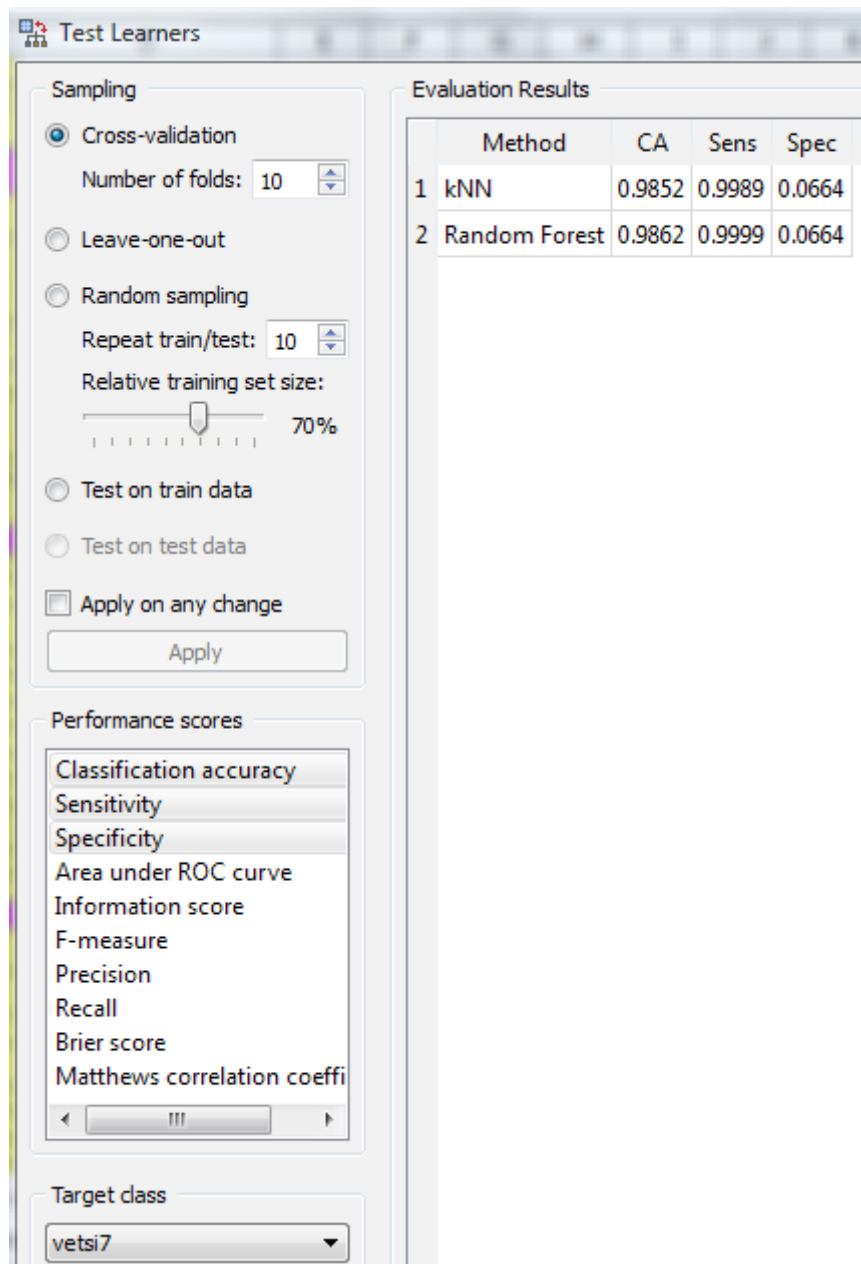


Figure 23: Example of evaluation of the result from Chapter 6 in Orange software. Classification target is annotation “ANOTACE_apgar_suma_5” with class “vetsi7”. It shows classification accuracy, sensitivity and specificity of Nearest neighbour (kNN) and Random Forest classifiers. 10-fold cross validation is used.