Selected Data Mining Methods and Their Applicability to the Television Audience Monitoring Data in the Czech Republic

Jan Walter
Department of Logic
Faculty of Philosophy and Arts, Charles University, Prague
j@n.cz

Advisor: Prof. RNDr. Petr Hájek, DrSc.

September 11, 2006

Declaration of unaided work

Hereby I declare that this thesis is my own unaided work and that I used sources mentioned in References exclusively.
Prohlašuji, že jsem tuto práci vypracoval samostatně, a použil výhradně citovaných pramenů.

Contents

Abstract

Data mining is nowadays a fast-growing field, which incorporates machine learning, statistics, and logic within computer science. It has the potential to bring new insights into almost all branches of human activity, because the data are stored almost everywhere. This thesis tries to show the main aspects of the original Czech method GUHA, to demonstrate its strength via its application to television audience data, and finally to compare it with the association rules method, which is similar to it. The ambition of this text is to interconnect the world of praxis with the theoretical field, where methods are invented. It also serves as an introduction to data mining itself. The results show that GUHA is a full-value method with several interesting features and might be a good tool for extracting knowledge from analyzed data.

Acknowledgement

I would like to express my deep and sincere gratitude to my advisor Professor Petr Hájek for the inspiration and methodical leadership he gave me, my wife Daniela for her boundless patience, people from the Department of Logic for the knowledge base they tried to pass me and their patience as well, ATO and Mediaresearch for the data they kindly provided me and Kevin for his language proofreading. Not forgetting several others for their support.

Notice

Source of all data published and analyzed: Electronic Measurement of Television Audience in the Czech Republic, ATO-Mediaresearch.

Chapter 1

Introduction

Together with a fast development of Information Technologies (IT) during the second half of 20^{th} century, database systems attached attention to themselves. They primarily serve to store mass data in a structured form to be subsequently retrieved. They soon became an integral part of all big companies, both governmental and non-governmental organizations as well as research centers. Consecutively diverse implementations spread among mid-sized and also small companies, because they enabled distinct simplification or even complete automation of routine processes. Thus they reduced cost and increased the quality of their products and services. Database systems together with the Internet brought synchronized sharing of information to users worldwide.

Plenty of such systems have been filled with data during last few decades, their functionality achieved a high quality in speed and stability. However, having these large data warehouses does not only imply to insert, update and read the data back in the same way they were stored. It also represents a huge base of observed cases, where new relations that cannot be seen at first sight can be explored. Such new knowledge might be of a greater value compared to initial raw data.

A fast developing discipline that deals with finding new knowledge inside collected data is called *Data Mining (DM)*, alternatively *Knowledge Discovery in Databases (KDD)*[1]. The first world-renowned papers and books about related techniques were published in early 90's. However, as it sometimes happens in the history of Czech lands, a group of Czech scientists developed their own approach more than 20 years earlier. Their effort ended up in a comprehensive book by *Hájek* and *Havránek* [16].

[1] Some authors like *Fayyad* [8] explain DM as a part of overall KDD process, but I do not consider it that important.

Nowadays, the term data mining encompasses a great amount of tools based on the theoretical background of statistics, logic, machine learning, artificial intelligence. Their typical representatives include association rules, decision trees, Bayesian belief networks, neuron networks, linear regression, cluster analysis, methods based on analogy and many, many others. Some of them have passed through a rich history and their soundness is verified by decades of usage, others have arisen recently and are still being improved.

This work aims to provide the reader who possesses the essential knowledge of mathematics, logic and database systems an overview of theoretical principles of contemporarily frequently used *association rules*, to compare them with the original Czech method GUHA *(General Unary Hypotheses Automaton)*, and subsequently to apply them to *television audience monitoring data* to make a comparison of their explorational and expressional strengths.

1.1 Neighboring disciplines

As we have already mentioned, data mining itself contains numerous methods, algorithms and other computational approaches originating in a wide area of diverse mathematical, machine learning and data management branches. Although most of them were designed for a specific range of purposes, they cay all serve in various kinds of situations.

Some of the most obvious scopes of usage belong to medicine and biophysics, weather and climate sciences, customer relationship management (CRM) and marketing in general, banking, image processing, etc. In general, data mining methods can be used anywhere, where data is measured, collected and saved in a form of quantified values.

In medicine, we can analyze what symptoms can possibly lead to a particular diagnosis. A typical database made for similar purposes contains the description of many thousands cases of patients collected during years to decades. Each patient is observed from many points of view throughout his medical history. A data mining task usually involves the following:

- capturing trends in changes of discrete attributes and continuous values in a time perspective,
- finding out, what symptoms (IE COMBINATION OF ATTRIBUTES AND VALUES) correlate with target diagnosis,
- combining recognized trends and correlations to build up a knowledge base that helps experts (IE DOCTORS) to identify the disease in its early stage and prevent its further progress.

Other examples lie in the field of customer-merchant relationship. Even the smallest shop usually stores cash desk transactions in a database. Thus we can retrieve what types of products tend to be sold together in a single basket (during one purchase). As credit cards stay the same for long time, we can also follow a single customer's history and analyze seasonal trends. This type of data mining task is called *market basket analysis*.

The knowledge of what commodities, brands and concrete products are bought jointly help the merchant to understand his client's priorities and interests. Then he can accordingly reposition the goods in the shop to sell the friendly products (EG BEER AND CHIPS) together and design an effective system of sales and product coupons to support less favorite ones. The adoption of loyalty cards associated with customer registrations can enrich the value of a transaction database in a other qualitative dimension that may help to build up a small product campaign in bigger stores.

Data mining methods can also be very helpful in banking and insurance. FOR INSTANCE, IF THE CLIENT ASKS FOR A LOAN, THE BANKER LETS HIM FILL IN A WIDE-RANGING QUESTIONNAIRE DETECTING MANY FACTS THAT COULD AFFECT PAYING OFF THE LOAN. THEN HE LETS THE *scoring function* BASED ON PROGRESS OF PREVIOUS CASES CALCULATE THE PROBABILITY THAT THIS CLIENT WILL BE A GOOD AND PROSPECTIVE NEW AND CATEGORIZE IT INTO RISK GROUPS. IT CAN MEAN IN CERTAIN IMPLEMENTATION THAT RISKIER CASES HAVE TO PAY GREATER INTEREST TO BALANCE THE POSSIBILITY THAT THEY DEFAULT.

In general, merchants use data mining tasks to:

- segment clients to understand their behavior, priorities and interests better,
- recognize behavior deviations to minimize clients loss,
- develop interesting campaigns and programmes to maximize client gain and measurable popularity,
- monitor the influences of competitors' moves on their profits.

We have roughly described some of the main occurrences of data mining tasks in a few distinct branches. The list would never be complete because the discussed methods can be used almost anywhere. However, there are several interesting spheres of application we should mention to make the notion of this broad interdisciplinary subject more colorful:

- Sound and image processing, which include voice and character recognition or autonomous navigation systems.

- Text mining that helps to understand the content of text documents in terms of their classification, querying, and comparison. Typical implementations include internet search engines or spam filters.
- Sociohistorical implementations trying to justify reasons for outstanding sociological changes in the past (EG WARS, MIGRATION, ECONOMIC RECESSION).

1.2 Data mining complexity

As we have seen in previous paragraphs, data mining represents a broad and very complex area of applications. Each of them with its autonomous history indicates a huge amount of approaches that participated in the constitution of this discipline. Although we are now still in the beginning stage, we can take advantage of lots of work that has already been done in this field.

Thanks to that (and maybe just because of that), we shall consider every new data mining task prudently. A good selection of used methods makes more than a half of the given task done. Conversely, if random tools are used to extract some knowledge from laboriously gathered observational data, one can easily get results, which, if deployed, would lead to illusory interpretations.

The first part of every such job constitutes a good introduction to the investigated domain. It is necessary to state clear aims. Understanding the data and its structure is the next important step to come. FOR INSTANCE PARTICULAR VARIABLES CAN BE *discrete* OR *continuous*, WHICH IMPLIES DIFFERENT APPROACHES TO THEM. MANY METHODS WOULD COLLAPSE IF THIS DIVERSITY WAS OMITTED.

Once we understand the principles, processes and important moments of the field and know the structure of the data, we can advance to work with data itself. Firstly, it may be necessary to format the data into a form that fits the used tools. Than we can process them and interpret their outputs. However, this is not the end. Most algorithms and methods can be variously parameterized to suit different situations. It is usually necessary to re-run them with different settings repeatedly to get meaningfully interpretable results. Only the comparison of outputs of different methods considering their advantages and drawbacks leads to desired conclusions. One such general methodology is described in section 2.3 thoroughly.

1.3 Motivation of this paper

Data are stored in databases all over the world. They are related to almost any human activity, and their growth often reaches size of terabytes. For a long period of time, it has not been possible for experts to manually explore

new knowledge patterns inside these large data sets. Instead, they have to use different automated methods to filter out possibly interesting coherence. However, the interpretation and further usage of automatically found relations still remains their job.

Data mining tasks can be, therefore, formulated in many diverse spheres and situations including the non-commercial sector. The discussed methods can help mankind to solve problems concerning health, genetics, cosmology, nature, history and thus push our wisdom a big step forward. And this fascinates me. GUHA is a rather rich and well defined system for exploring observational data. Despite its age, it can be compared to contemporary techniques. I would like to enlighten as to its principles and contribute to its wider usage.

This paper contains a basic introduction to data mining, its methods and algorithms, their classification and way of usage. It concentrates on the original Czech method GUHA and other comparable techniques and their usability with television audience monitoring data. It, in fact, builds upon a group of concrete data mining tasks by theorizing on the usability of different methods in this specific, but not unique situation. A certain criterion for the choice of reviewed techniques arises from this claim: they have to bring a similar type of knowledge about source data as GUHA. Others will only be mentioned.

1.4 Typographical convention

To make the document more readable, I will distinguish examples and practical usage descriptions with SMALL CAPITALS, and new terms, key words and indexed notions with *italics*.

Chapter 2

Theoretical Basis

2.1 Introduction to data mining

Hand, Mannila and *Smyth* [17] provide at the beginning of their text the following informal definition:

> Data mining is the analysis of (often very large) observational data sets to find unsuspected relationships and to summarize the data in novel ways that are both understandable and useful to the data owner.

Requirements mentioned above are very really essential. What do they mean? The *novelty* of extracted information constitutes a cornerstone for everything we would call data mining. We want it to bring us new (or *informative* as we call it in logic) knowledge that we did not already have, which is closely connected to its *interestingness*. For instance, if we got the relationship "Everyone who is pregnant is female", we probably would not consider it as very surprising and therefore it would not be considered a valuable result of the data mining task. At the moment unfortunately, just a few algorithms try to take into account prior (much less background) knowledge at the moment.

Reported outcomes also have to be *understandable*, which means not only legible by humans but also as simple as possible to interpret. A set of cubic equations on quantitative variables of our data set (see later in this chapter) might present some new knowledge but no mortal would understand its actual message. Only the combination of interestingness and understandability of discovered relationships make the results of the data mining procedure *usable*, because they are valuable and interpretable, and therefore can be utilized to improve processes underlying the generation of source data.

Another, rather philosophical question related to the discussed topics comes up: *"Can the data mining procedure be automated, or is some interaction with an expert needed?"* This is strongly connected to background knowledge, because algorithms can hardly distinguish between *correlation* and *causality*, while the user is interested in causality primarily. If the final report included the relationship between birth rate in China and amount of cabbage planted in New Zealand, no one would ever believe its interconnection. We will leave the question open; in any case, we cannot nowadays rely on algorithms boundlessly in this way. An expert has to be present at each data mining task to sort extracted information according to its value. On the other hand, routine processes with strictly a defined form of input data as well as output reports can certainly run autonomously.

2.1.1 What is data?

As it arises from the term itself, data mining relates to some processes working with *data*. But what does the word itself usually mean? Typical data mining algorithms and techniques treat data in a form of matrix. It consists of rows corresponding to single observed *cases* (eg measurements, respondents, transactions, etc.) and columns representing *variables*. They are sometimes, depending on context, called *dimensions* or *attributes*. Of course, the raw source data often do not completely fit in the matrix-form framework, however most contemporary algorithms expect it in this joined manner.

Variables of the data set can be divided into following basic groups:

- *Quantitative* ones that represent *continuous* quantities.
- *Categorical* ones that represent *discrete* quantities. This means that the finite set of values it can get is known in advance. These variables are
  - either *nominal*,
  - or *ordinal*

  depending on, whether the values can be ordered in a natural way. If it has its reasonable interpretation (eg Higher Education < Basic Education), the variable is ordinal and enables additional computations (eg average, linear regression) with numeric values assigned to labels (eg age spent in schools).

Numeric values of variables may, as an addition to pure numbers, represent diverse scales (time, spatial position), which brings to data mining tasks new dimension and expands the *hypothesis space*, the area where new, possibly interesting, knowledge is sought.

As well as with almost everything in the life, the data are not always perfect. Several typical discrepancies can be often found in the source data entered into data mining procedures. They are usually *noisy*, meaning that they contain *errors*. It arises at the moment when reality differs from a measured value. This deviation can be caused, for example by an untrue answer of a respondent, an imperfect device etc. Raw measured data are usually *cleaned* but this process does not necessarily have to repair the data completely. Such errors may lead to biased estimations, and furthermore to inner contradictions. Some basic algorithms assume data without errors, however, those employed in real situations have to be robust enough to be prepared for this possibility.

A slightly different situation emerges at the moment when we completely miss some measurement. *Missing values* introduce another problem in devaluating the data. Although we do not know the actual value, we at least know this fact. It typically happens when the respondent refuses to answer some question, or when we do not have the opportunity to measure a quantity in some case. This is different comparing to errors. The simplest intuitive approach is to exclude a case with such a property from further computations, however, this might be waste of data, if it is well measured in other attributes. Several methods solving this situation (both in general and inside particular algorithms) are used to overcome this possibility.

2.1.2 Main features of algorithms

Every data mining technique is, above all, determined by its

- primary *purpose*, which usually means to:
  - *View* the data in some aggregated, possibly graphical form.
  - *Find consequences* in the form of patterns, catching relationships inside segments of data or
  - *understand* the data as a whole, which means to describe their generator, in other words to model the process of their growth.
  - *Predict* some aspects of the future or of cases to come thanks to previously described modeling.

  and
- top-level *components* that enumerate:
  - *Target structure*, the way the discovered knowledge is represented.
  - *Score function*, the way, in which the quality of the target structure is evaluated by.
  - *Optimization method*, the way of achieving the best performance of the target structure in terms of score function.

The aspects mentioned above, mainly those related to the purpose of the task, build the base for the taxonomy of data mining algorithms as more closely described in section 2.2.

2.1.3 Common bottlenecks

In data mining, two main demands appear and they usually tend to go against each other. Firstly, we want to reach the maximal accuracy possible, so that the extracted knowledge really fits the data its based on. On the other hand, we desire the new knowledge to be valid and thus usable in similar situations, with analogous, however not exactly the same, data (IE FOR CLASSIFICATION OF FUTURE CASES.) If the extracted knowledge gets too complex and describes the analyzed data too closely, it loses its universality. This is called *overfitting* the data set (see Figure 2.1 for a demonstration of this). The problem can be solved by diverse techniques in different situations, some algorithms already have such an approach built-in, on the other hand, there are a few methods usable in general. We will mention a few of them.

- There are at least two similar approaches that prevent the target structure from being too complex internally. One of them *stops its generation* before it is perfect for the source data, the second one *simplifies* it afterwards (EG THE PRUNING OF TREES OR NEURAL NETWORKS).
- Another way that helps to estimate proper target complexity is based on the division of source data (which may be either random, or stratified). One part (the *training set*) is used for learning and the other one (the *test set*) for testing. This *cross-validation* brings us information about the performance on a different data set, than the knowledge was based on. If it is made just once, it still does not have to reduce overfitting significantly. If we repeat the division iteratively we get very robust estimation. The procedure is executed n times, each time the data of size S are divided into two groups, where S/n cases represent test data and the rest is uses for training. One case is therefore used once for evaluation and $n-1$ times for learning. In every step, the complexity C of the target structure (EG THE NUMBER OF EPOCHES IN A NEURAL NETWORK, THE SIZE OF THE DECISION TREE, ETC.) with best performance on the validation set (see Figure 2.1) is remembered. Then we get its average $\bar{C}$, which makes the target complexity of the final model learned on the complete data set.
- Somewhere between the fashions depicted above lies the *minimum description length (MDL) principle*. It rises from a hypothesis with a very long history (which comes from renaissance philosopher William of Occam and often is referred to as *Occam's razor*), which states that *the simplest hypothesis fitting the data is preferred*. MDL is a measure based on the *information description length* of the model and exceptional cases together - summary representation of the data. This means that both the size of the model and number of cases not described by the model contribute to the

description of its quality. Thus the simpler model may be less accurate but still preferred. It results into a lower probability of overfitting.

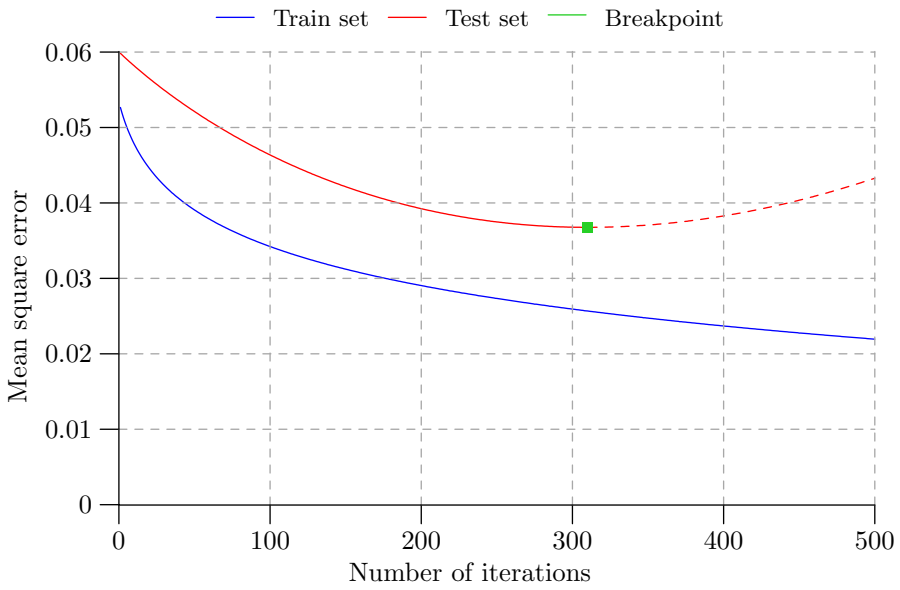

Figure 2.1: Overfitting demonstrated with the example of iterations (weight updates) of a neural network. The blue line shows the permanently improving performance of a model upon the training data during epoches, while at a certain point, the performance upon the validation set begins to get worse.

Another common risk in data mining is often named as the *curse of dimensionality*. Treated data often have many dimensions reaching above the hundreds, which implies two main constraints:

- Distances between cases in multidimensional space grow. We know that 95% of cases lie within a confidence interval $[\mu - 1.96\sigma; \mu + 1.96\sigma]$, where μ is the mean and σ is the standard deviation, for a normally distributed variable. If dimensionality increases to $d = 40$, about 10% of the data occur within the distance 1.96σ from the center and if d goes above the hundreds, there are practically no data around the center. Therefore, algorithms based on this measure hardly outperform.
- As the number of dimensions increases, the size of the hypothesis space grows as well. Since a set S has $2^{|S|}$ subsets, the growth is exponential, which makes the exploration computationally very complex.

The curse of dimensionality is usually overcome by weighting the importance of particular dimensions. Not all dimensions play the same role in the data, their

variables have different variance etc. Different methods (EG PRINCIPAL FACTOR ANALYSIS OR FACTOR ANALYSIS) can be employed to *reduce the dimensionality*, or simply just a few dimensions can be chosen to be involved in the process of knowledge extraction.

2.2 Taxonomy of algorithms

Data mining algorithms and techniques can be classified according to tasks they accomplish. The particular implementation of components (EG THE NAME OF A TREE BUILDING ALGORITHM, PROGRAMMING LANGUAGE, DATABASE MANAGEMENT SYSTEM, ETC.) is not of the topmost importance, however its aim is a primary concern. One substantive comment should be made before we state individual categories: their borders are not sharp. Some methods lie between them, or can be used for more purposes depending either on which outputs are employed or on the way of usage. They also sometimes interact between themselves.

The basic analyses with added value that can be done with data sets concern *summarization and visualization*. It is very hard to capture the meaning of more than 10 rows of a data set, abstracting some knowledge from a hundred rows is impossible for human mind. Methods such as *pivot tables* simplify the view of data in such a fashion that a human can essentially understand what is going on in the data essentially. They enable the user to look at them from a greater distance. Graphical representation of this condensed form makes it even more usable. There is a broad spectrum of charts that visualize data, ranging from simple pie charts and 2D graphs to methods that geometrically project huge multidimensional data sets into 3D space. The visualization of outputs of other data mining methods is also a great benefit.

If we go further in mining the data, we find connections in the form of *patterns* and *rules*. They represent relationships among some variables inside particular groups of cases that can be relatively small in comparison to a whole data set. FOR INSTANCE, IN BANK CREDIT DATA, WE CAN FIND A RULE THAT STATES "YOUNG PEOPLE TEND TO DEFAULT IN SUMMER MONTHS MORE OFTEN THAN OTHERS". EVEN IF THE DATA HAD HUNDREDS OF COLUMNS, THIS RULE CONNECTS ONLY AGE AND FREQUENCY OF DEFAULTS AND REFERS TO A SUBSECTION OF THE DATA COLLECTED IN SUMMER. Patterns and rules often help us in finding outlying anomalies, small portions of the data that prove qualities significantly distinct from the rest. *Association rules* described firstly by *Agrawal et al.* in [2] represent the very classical example of such approach. They and other related methods will be described more closely in section 2.4.

On the other hand, there are methods that cover a complete input data set. This means that they try, in some manner, to describe the data as a whole, to explain their overall nature or each distinct component. Such descriptions,

or *models*, which are used in this context more often, typically include *concept learning* methods, the *probability distribution* of values leading to *density estimates*, partitioning techniques usually called *segmentation*, which is rather a marketing term, or, from a technical point of view, *cluster analysis*. Besides the aforementioned, we must not forget GUHA.

As we manage to learn new knowledge about our data and understand their inner principles, we can better approach an estimation of future cases better. WE MAY, FOR INSTANCE, TRY TO PREDICT TEMPERATURES BASED ON PAST OBSERVATIONS AND MEASUREMENTS OF WEATHER AND CLIMATE, OR THE POSSIBILITY THAT A CLIENT OF A BANK WILL DEFAULT. *Regression* and *classification* methods handle similar situations most often.

Especially with the recent development of the internet and digital multimedia, a new strong branch of stand-alone data mining methods arises. *Content retrieval* techniques serve to search huge storage places containing texts (VIZ WEB SEARCH ENGINES) and even more complicated audio-visual material to find similar, or related objects, and patterns inside them.

2.3 CRISP-DM

Before we advance to the description of compared methods and their application, we should mention an interesting attempt to standardize the treatment of every data mining task. Such an approach was developed by the CRISP-DM (CRoss-Industry Standard Process for Data Mining), a consortium of originally three companies experienced in the field of data mining (DaimlerChrysler, SPSS and NCR) and its Special Interest Group having over 200 members. The resulting document [7] from summer 2000 defines an abstract and general data mining methodology, that is even prepared for new and unforeseen situations and techniques.

Its model is based on the hierarchical assignment of responsibility to individual levels *(phases, generic and specialized tasks, process instances)* of the overall process that takes place in different contexts depending on *application domain*, *data mining problem type*, *technical aspects* and used *tool and technique*.

Since the idea of reasonable standardization is very close to my attitude, I will try to follow the main guidelines of CRISP-DM methodology in the application of data mining methods to television audience data in chapter 4.

2.3.1 The CRISP-DM reference model

The life cycle of each data mining project is made of 6 phases. Together with mutual relationships they build up the *CRISP-DM reference model* as depicted

in Figure 2.2. Arrows demonstrate strongest dependencies, thus connections between the output of one phase and the input of another.

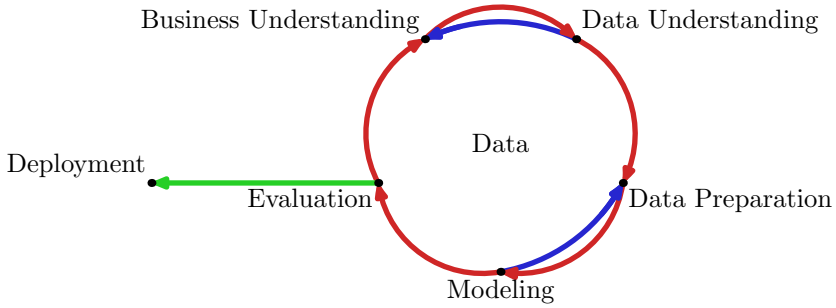

Figure 2.2: Top level phases of the CRISP-DM process

Business understanding The initial task is to determine the main goals of the project that should be accomplished in terms of business understanding, taking into account also outer connections and influences (EG THE POSITION ON THE MARKET, COMPETITORS, ETC.). The definition of success of the project is also a very important point in the first phase. Another task is related to the logistics of the project, which means to assess available resources, including people with knowledge, hardware and software, and consider possible risks and constraints that may influence its progress.
Then we can state the project's data mining goal in its technical details and outline its plan within this context, together with a selection of used methods and tools.

Data understanding The cornerstone of the second phase is the localization of data in all relevant resources and their collection. A centralized source data set is subsequently analyzed to recognize its format, explored to find out essential properties (VIA SIMPLE STATISTICAL ANALYSES, EG DISTRIBUTION OF VALUES) and described for further usage. In this moment, the quality of the data, including its completeness, correctness, level of noise, missing values and possible solutions of these problems, is evaluated as well.

Data preparation Data preparation is a necessary step to get the data in a proper form for input into mining procedures. It means to select the minimal subset needed to obtain the demanded results, clean it (OR USE ANOTHER METHOD LIKE ESTIMATION OF MISSING VALUES) to minimize the influence of factors that decrease quality of outputs, derive new attributes from original ones, and possibly add new records representing those not contained in the source data set. Depending on a concrete used tool, it is often necessary to convert the values of individual variables into proper syntactic format.

Modelling This part of the process stands for the selection of methods to extract new knowledge, building a desired model or whatever the task is. Before we apply them to prepared data, we have to design a way to test the quality of its results, which may sometimes mean some additional data preparation like when using cross-validation. Then we adjust parameters of used techniques and run procedures, if needed, repeatedly. Results are assessed and, optionally, other approaches and methods are employed for comparison. The best models advance to the next phase.

Evaluation As we acquire new evaluated models or other generalized knowledge, we can compare their message with the original business task, review the whole process reflecting new experience, and consider the following steps. According to their adequacy, we either return to the beginning with new experience and knowledge, or finalize the complete procedure.

Deployment The purpose of the last phase is to utilize the previous procedure in reality. Depending on the business task, it may imply simple reporting, or sophisticated implementation of repetitive data mining processes, as well as changes in the business strategy of company involved.
The evaluation of the overall process and conclusion comprise the end of the entire project.

2.4 Compared methods

In this chapter, we will describe several applied techniques in detail. The description of each technique will include background information, taxonomy subsumption, and a thorough explanation of used algorithms and their parameters, including a theoretical justification mentioning possible alternatives and improvements.

It is my intension is to explain to the reader the main principles of applied methods in depth, however, because they will be used in chapter 4 and a good understanding of them is necessary for a well-founded comparison, only the essential or most often used mutations will be discussed. Many alternative approaches and ways of improvements of implementation (especially towards space/time efficiency) may be found in additional literature. Some of the most important articles related to this topic are mentioned in section 2.6.

2.4.1 Association rules

An *association rule* is undoubtedly one of most typical ways of representing the knowledge discovered in databases. It describes connections between individual variables and their values. In general, its basic form is $A \Rightarrow C$, where A is an *antecedent* and C a *consequent* of the rule. Both the antecedent and the consequent are conjunctions of atoms that define allowed value for a categorical

variable. Although classical examples of association rules refer to binary variables (eg a basket containinig/not containing specified commodity), we can handle it in this, more general way, allowing a finite number of values. The rule itself has some (usually 2) additional parameters that determine its:

- strength or interestingness and
- statistical as well as business significance.

Let's have a look at a very classical textbook example of association rules usage. It is an analysis of the content of shopping baskets. We mainly refer to [2], [1] and [3]. Thanks to bar code technology and its connection to cash registers in shops, we gain a large database of transactions, which represent a single basket. This means that we know which commodities were bought together, or even by one customer anytime in history, if we identify them somehow (for instance via credit card numbers). We can then, for example, get a rule stating: 85% of people who buy fish, also put frozen chips into their baskets, and these comprised 8% of all customers (support and confidence approach, see next paragraph). As already mentioned in chapter 1, knowledge of this type may be very helpful to marketers in many ways.

The most frequent form of association rules has two parameters, *support* and *confidence* defined as:

$$\begin{aligned} sup(A \Rightarrow C) &= \frac{|D_{A\&C}|}{|D|} \\ conf(A \Rightarrow C) &= \frac{|D_{A\&C}|}{|D_A|} \end{aligned}$$

where D_φ stands for a set of objects in the database satisfying condition φ and $|D_\varphi|$ is its size; $D = D_\top$. Support is a proportion of objects satisfying both antecedent and consequent to all objects in the database, therefore something like statistical and business significance of the rule. Confidence says what portion of those satisfying A also satisfy C.

The essential algorithm finding rules with minimal support and confidence is called ***Apriori***. It stands somewhere between brute the force testing of all possibilities, which is exponentially complex, and newer, more space and time effective heuristic alternatives referred to in section 2.6. Its principles were formulated in [2] and [3]. The algorithm is based on the simple observation that sets of items satisfying some condition cannot grow with a stronger condition. It is, then, divided into two phases. Firstly, a set of *conjunctions of potential antecedents and consequents*, thus candidates for rules with minimal support is generated. The term *large itemsets* is often used in this context in literature. Nonempty conjunctions A of these candidate conditions P are subsequently step-by-step tested for confidence of a rule $A \Rightarrow P \setminus A$.

```
1: /* number of variables (maximal length of conjunction) */
2: k = |Variables|
3: res = ∅
4:
5: /* initialization */
6: for (i = 1; i ≤ k; i++) do
7:    P_k = ∅ /* system of potential candidate sets */
8:    C_k = ∅ /* system of candidate sets */
9: end for
10:
11: /* get all sets of candidate conditions */
12: for (i = 1; i ≤ k; i++) do
13:    if i=1 then
14:       P_i = Atoms
15:    else
16:       /* extend to conditions of size i */
17:       P_i = generatePotentialCandidates(i, C_{i−1})
18:    end if
19:    if P_i = ∅ then
20:       break
21:    end if
22:    for all p ∈ P_i do /* test for minsup */
23:       if |D_{∧p}|/|D| ≥ minsup then
24:          C_i = C_i ∪ {p}
25:       end if
26:    end for
27: end for
28:
29: /* examine candidates for rules */
30: for (i = 2; i ≤ k; i++) do
31:    if C_i = ∅ then
32:       break
33:    end if
34:    for all cond ∈ C_i do
35:       conseq = {{x}; x ∈ cond}
36:       /* initial call w/ atoms as consequents */
37:       res = res ∪ generateRules(i, 1, cond, conseq)
38:    end for
39: end for
40:
41: return res
```

Algorithm 2.1: Apriori

```
1: /* This sub-procedure generates new potential conditions of size i from can-
   didates of size i − 1 */
2: procedure generatePotentialCandidates(i, Prev)
3:
4: /* initialize candidate set to be returned */
5: res = ∅
6:
7: /* extend Prev conditions to length i mutual combinations */
8: /* use all candidate pairs of size i − 1 */
9: for all x ≠ y; x, y ∈ Prev do
10:   /* we assume that atoms are ordered in x and y */
11:   if x_{i−1} = y_{i−1} then
12:     continue /* get next pair */
13:   end if
14:   for (j = 1; j ≤ i − 2; j++) do
15:     if x_j ≠ y_j then
16:       continue 2 /* get next pair */
17:     end if
18:   end for
19:   /* add {x_1, ..., x_{i−1}, y_{i−1}} */
20:   res = res ∪ {x ∪ y}
21: end for
22:
23: /* check presence of all (i − 1)-long sub-conditions in all candidates to be
   returned */
24: for all c ∈ res do
25:   for all sub_c ⊂ c, |sub_c| = i − 1 do
26:     if sub_c ∉ Prev then
27:       res = res \ {c}
28:     end if
29:   end for
30: end for
31:
32: return res
```

Algorithm 2.2: Generate potential candidates

```
1:  /* This sub-procedure generates rules from candidate sets */
2:  procedure generateRule(size, i, Cond, Conseq)
3:
4:  res = ∅
5:
6:  for all (seq ∈ Conseq) do /* test for confidence */
7:     conf = |D_{∧Cond}|/|D_{∧(Cond\seq)}|
8:     if (conf ≥ minconf) then
9:        res = res ∪ {Cond \ seq ⇒ seq}
10:    else
11:       Conseq = Conseq \ {seq}
12:    end if
13: end for
14:
15: /* possibility to move one more atom to consequent */
16: if size > i + 1 then
17:    Next = generatePotentialCandidate(i + 1, Conseq)
18:    res = res ∪ generateRules(size, i + 1, Cond, Next)
19: end if
20:
21: return res
```

Algorithm 2.3: Generate rules from candidate sets

Now, we can formalize the complete `Apriori` (see algorithm 2.1) and explain its main principles more deeply, if needed. Atoms made of categorical Variables and their values are building blocks of conditions (temporarily stored in helper sets P_k and C_k). Parameters minsup and minconf define lower bounds for rules we are interested in. The `Apriori` algorithm uses two subprocedures *generatePotentialCandidates* (algorithm 2.2) and *generateRules* (algorithm 2.3).

After initialization the first step is carried out. It iterates over number of variables beginning with Atoms (line 14) and continuing with conditions made of shorter ones (line 17). In each cycle, it tests potential candidates for the rule base (*antecedent* & *consequent*) whether they satisfy minimal support. If they do, they are included in the set of candidates C_i (line 24). The *generatePotentialCandidates* call (line 17) returns base conjunction sets of length i based on those in C_{i-1} of length $i-1$, which satisfy minimal support, because a stronger condition cannot satisfy it without being satisfied by its sub-condition.

The second step verifies rules that can be made of the base conditions. The candidate set containing 3 conditions $\{A, B, C\}$ may form 12 rules (see Figure 2.3).

length of base condition (i)			
2		3	
$A \Rightarrow B$	$B \Rightarrow A$	$A\&B \Rightarrow C$	$C \Rightarrow A\&B$
$A \Rightarrow C$	$C \Rightarrow A$	$A\&C \Rightarrow B$	$B \Rightarrow A\&C$
$B \Rightarrow C$	$C \Rightarrow B$	$B\&C \Rightarrow A$	$A \Rightarrow B\&C$

Figure 2.3: 3-element set may produce 12 rules

If we remember the formula for the computation of confidence, we realize, that if $A \Rightarrow B\&C$ satisfies the minimum, $A\&B \Rightarrow C$ satisfies it as well. We, therefore, begin testing rules with atomic consequents and in each step test rules with longer consequents. The subsequent iteration uses only those consequents generated via *generatePotentialCandidates* (algorithm 2.2) run on previously verified consequents. $A \Rightarrow B\&C$ is tested only if both $A\&B \Rightarrow C$ and $A\&C \Rightarrow B$ hold.

generatePotentialCandidates firstly combines all pairs that differ only in the last position (assuming common order) by adding their union (line 20) among potential candidates (this is a slightly improved version of combination of all pairs that differ in one point, not taking into account their order). Then, an additional test for a necessary presence of all conditions shorter by one in original candidate set C_{i-1} is made (line 26).

generateRules tests all rules of form $Cond \setminus seq \Rightarrow seq$ (line 8) for confidence. If it is satisfied, the rule is included in the result, otherwise the *seq* is deleted from the set of potential consequents. If we can move at least one more condition from antecedent to consequent, we generate next set of candidates for consequents and *generateRules* for them (lines 17-18).

2.4.2 GUHA

Guha (General Unary Hypotheses Automaton) is a method that originated in Bohemia at the end of the 1960's. Its basic principles were formulated in a series of articles in Czech [12], [14], [15] and English [13] that resulted in [16] (the complete history of Guha is discussed in [11]). The very first hypothesis found by computer implementation stated that pregnant epileptics are women.

It is based on the 1^{st} order two valued logic with unary predicates, in which matrix data form a model of the theory. Rows are objects and $A_1, ..., A_n$ their attributes. Each attribute has its finite domain $Dom(A_i)$, a set of admitted values. $P_{i,j}(x)$; $x \in D$ says that data object x (represented by a row in the matrix) has property $P_{i,j}$, thus its value in attribute A_i is $A_{i,j}$. Without loss of generality, we can assume $\forall i |Dom(A_i)| = 2$. Domains that accept only 1 value are not interesting and others can be equivalently encoded in binary. It produces a matrix of 0s and 1s. Objects satisfy (the row x of the matrix has

1 in i^{th} column) or do not (0 in i^{th} column) the atomic formula $P_i(x)$. Since we use just one variable in formulas, we can omit its writing and use atomic formulas $P_1, ..., P_n$. Common logic connectives (&, $\vee$, $\neg$, $\rightarrow$) are used to form open (the variable they refer to is not bound by a quantifier) formulas LIKE $(P_1 \& \neg P_2) \rightarrow (P_3 \vee P_4)$. Each open formula φ and a row r of the data matrix is evaluated a truth value $||\varphi||_D[r]$ 0 ($false$) or 1 ($true$) in a standard fashion. $Fr_D(\varphi) = |D_\varphi|$ (D_φ means the same as in previous subsection 2.4.1, it stands for a number of objects satisfying open formula φ. The interpretation of quantifiers is also made normally as shown in Figure 2.4.

quantifier	formula	condition to be satisfied	
$\forall$	$(\forall x)\varphi$	for all rows r from data D $\|\|(\forall x)\varphi\|\|_D[r] = 1$	$Fr_D(\varphi) = \|D\|$
$\exists$	$(\exists x)\varphi$	for at least one row r from data D $\|\|(\exists x)\varphi\|\|_D[r] = 1$	$Fr_D(\varphi) \geq 1$

Figure 2.4: Interpretation of classical quantifiers in GUHA

Extending it, GUHA comes with a generalized view of quantifiers. Formula $\forall x\varphi$ with the classical "for all" quantifier demands all objects to satisfy φ to make the whole formula true. As we live in a colorful world where different shares of the whole may be interesting in a variety of situations, we do not always accept only conditions fulfilled by all objects. Newly introduced quantifiers are given by their arity (k, the number of sub-formulas they take into account) and truth function dependent upon 2^k frequencies (numbers of objects satisfied by logic combinations of single sub-formulas) and additional optional threshold parameters. THE SIMPLEST SUCH QUANTIFIER COULD BE CALLED *majority* AND MARKED BY $\surd_{50\%}$. $\surd_{50\%}x\varphi$ IS THEN TRUE IN THE DATA, IFF $Fr_D(\varphi) \geq \frac{|D|}{2}$ (AT LEAST HALF OF OBJECTS SATISFY φ). Classical quantifiers by definition conform this extension.

The strongest and also most useful group of quantifiers developed over years and included in GUHA is made of *binary quantifiers* of a form $A \Rrightarrow_T S$, where A is an antecedent (as in association rule), S a succedent and $\Rrightarrow_T$ the quantifier with T representing potential threshold parameters. Their truth function $Tr_{\Rrightarrow_T}(a, b, c, d)$ is computed from *four-fold pivot table* parameters as shown in Figure 2.5. $b = Fr(A \& \neg S)$.

		succedent	
		S	$\neg S$
antecedent	A	a	b
	$\neg A$	c	d

Figure 2.5: Four-fold pivot table for $A \Rrightarrow S$

Around twenty different quantifiers have been defined over years with together their particular properties that influence ways of usage. We will mention the most important of them and categorize them into classes that have a theoretical foundation and depend on properties of their truth function. *Implicational quantifiers* form a group with a very simple constraint:

$$Tr_{\sim}(a,b,c,d) = 1 \,\&\, a' \geq a \,\&\, b' \leq b \rightarrow Tr_{\sim}(a',b',c',d') = 1 \tag{2.1}$$

Its preservation condition is stronger than to require non-decreasing $\frac{a'}{a'+b'} \geq \frac{a}{a+b}$, so, for instance, the classical association rule based on confidence would fit it (not to mention support). This class might be expressed by words like *"many A's are S's"*. *Associational quantifiers* (aka *equivalence quantifiers*) form a superset of implicational quantifiers and is given by the following constraint:

$$Tr_{\sim}(a,b,c,d) = 1 \,\&\, a' \geq a \,\&\, b' \leq b \,\&\, c' \leq c \,\&\, d' \geq d \rightarrow Tr_{\sim}(a',b',c',d') = 1 \tag{2.2}$$

Double implicational quantifiers lie on a halfway:

$$Tr_{\sim}(a,b,c,d) = 1 \,\&\, a' \geq a \,\&\, b' \leq b \,\&\, c' \leq c \rightarrow Tr_{\sim}(a',b',c',d') = 1 \tag{2.3}$$

Another subset of associational quantifiers are *symmetric quantifiers* whose truth function fulfills:

$$Tr_{\sim}(a,b,c,d) \rightarrow Tr_{\sim}(a,c,b,d) \tag{2.4}$$

Comparative quantifiers as a subset of symmetric quantifiers require

$$Tr_{\sim}(a,b,c,d) = 1 \rightarrow ad > bc \tag{2.5}$$

$$(\text{for } a,b,c,d \geq 0 \text{ and } a+b>0\text{: } ad > bc \equiv \tfrac{a}{a+b} > \tfrac{a+c}{a+b+c+d})$$

thus S is more frequent in A than in whole data set.

The following table introduces the definitions of essential quantifiers together with their categorization. p stands for a parameter of each quantifier.

name	class membership [1]	truth function $Tr_{\sim_p}(a,b,c,d)$
founded equivalence	AS	$\frac{a+d}{a+b+c+d} \geq p$
double founded implication	ADS	$\frac{a}{a+b+c} \geq p$
founded implication	ADI	$\frac{a}{a+b} \geq p$
above average	AS(C)[2]	$\frac{a}{a+c} \geq (p+1)\frac{a+b}{a+b+c+d}$

Figure 2.6: Essential quantifiers

[1] [A]ssociational (Equation 2.2), [D]ouble implicational (Equation 2.3), [I]mplicational (Equation 2.1), [S]ymmetric (Equation 2.4), [C]omparative (Equation 2.5).

[2] For $p = 0$.

Binary quantifiers may be extended to the tertiary of a form $A \Rrightarrow_T S/C$, where additional C is a global condition that all objects referred to by the rule must satisfy. In other words, binary $A \Rrightarrow_T S$ is partialized to $D \restriction C$. Theoretical results in [10] show that some tertiary quantifiers (EG IMPLICATIONAL $A \Rrightarrow S/C \equiv A\&C \Rrightarrow S$) can and others (EG SATURABLE ASSOCIATIONAL) can not be transformed into a binary quantifier.

In addition to having a much broader theoretical background overlapping mainly to logic, which deals with classes of quantifiers, computational complexity and many other topics, implementations of GUHA procedures also test hypotheses with negation and disjunction of atoms with common property in both antecedent and succedent. THEREFORE RULES AS $Education[University \vee Secondary]\&Social\_classification[A \vee B \vee C] \Rrightarrow Salary[High \vee Above\_average]$ MAY BE FOUND. This aspect seems to be the greatest advantage comparing to association rules in practical usage.

2.5 Missing values

The handling of *missing values* is an important moment in the data mining task, as real data usually contain several variables with them. Different methods/implementations treat them in different ways, but we will concentrate on GUHA in *LISp-Miner* and the association rules in *Weka*. Table Figure 2.7 displays the initial frequency situation of cedents in the rule $\varphi \Rightarrow \psi$, where both formulas allow (might be evaluated) a *missing value* (neither *true* nor *false*) referred to as $m.v.$

$Fr(\varphi \Rightarrow \psi)$		ψ		
		$\top$	$\bot$	$m.v.$
φ	$\top$	a	b	e
	$\bot$	c	d	g
	$m.v.$	f	h	i

Figure 2.7: Initial 9-fold pivot table for cedents with missing values.

If the cedent ρ is atomic, $\rho(r) \equiv m.v.$ if and only if the variable ρ contains a missing value in the row r of the data. If the cedent is more complex, it happens when missing values are propagated from the atoms based on the following rules that are reminiscent of the behavior of logical connectives in Gödel's 3-valued logic. Their evaluation for object r is illustrated in the tables in Figure 2.8.

The following tables in Figure 2.9 depict 4 different strategies for the conversion of 9-fold (with frequencies a-i) table to a 4-fold table used by both association rules and GUHA. Their names help to understand their approach.

$\varphi\&\psi$		ψ		
		$\top$	$\bot$	$m.v.$
φ	$\top$	$\top$	$\bot$	$m.v.$
	$\bot$	$\bot$	$\bot$	$\bot$
	$m.v.$	$m.v.$	$\bot$	$m.v.$

φ	$\neg\varphi$
$\top$	$\bot$
$\bot$	$\top$
$m.v.$	$m.v.$

$\varphi \vee \psi$		ψ		
		$\top$	$\bot$	$m.v.$
φ	$\top$	$\top$	$\top$	$\top$
	$\bot$	$\top$	$\bot$	$m.v.$
	$m.v.$	$\top$	$m.v.$	$m.v.$

Figure 2.8: Propagation of evaluation of $\varphi(r), \psi(r)$ by $\top, \bot, m.v.$

non-deleting		
	ψ	$\neg\psi$
φ	a	$b+e$
$\neg\varphi$	$c+f$	$d+g+h+i$

deleting		
	ψ	$\neg\psi$
φ	a	b
$\neg\varphi$	c	d

secured		
	ψ	$\neg\psi$
φ	a	$b+e+h+i$
$\neg\varphi$	$c+f+g$	d

optimistic		
	ψ	$\neg\psi$
φ	$a+e+f+i$	b
$\neg\varphi$	c	$d+g+h$

Figure 2.9: 9-fold table conversion strategies

For each type of the quantifier (in terms of GUHA), we could build a hierarchy defined as

$$MVS_1 \triangleleft MVS_1 \equiv_{df} A \Rightarrow_{MVS_1} S \rightarrow A \Rightarrow_{MVS_2} S$$

where MVS_i stands for a *missing values frequency conversion strategy.*

IT IS, FOR EXAMPLE, OBVIOUS, THAT

$$MVS_{secured} \triangleleft MVS_{non-deleting} \triangleleft MVS_{deleting} \triangleleft MVS_{optimistic}$$

BECAUSE

$$\frac{a}{a+b+e+(h+i)} \leq \frac{a}{a+b+(e)} \leq \frac{a}{a+b} \leq \frac{a+(e+f+i)}{a+b+(e+f+i)}$$

AND

$$a \leq a+(e+f+i)$$

FOR ALL IMPLICATIONAL QUANTIFIERS, WHICH INCLUDE ASSOCIATION RULES BASED ON SUPPORT AND CONFIDENCE.

2.6 Further reading

Association rules represent a relatively simple and robust data mining method, but they also have some drawbacks. The most important seems to be the fact that it handles categorical variables only. [21] might be a good beginning; its approach is based on genetic algorithms. Improved implementations of the algorithm towards more effective ways were also published. Alternatives `AprioriTid` and `AprioriHybrid` are described also in [3], and [18] provides a systematic overview of other approaches. Another topic relates to ways to limit

quantity of found rules on large data sets and increase their quantity. In this context, other measures of interestingness of the rules besides confidence are introduced. The typical drawback of usage of confidence (minimal confidence holds even when antecedent and consequent are negatively correlated) and the proposal of its solution, are described, for instance, in [5], [24] and with an accent towards used algorithm in [27]. An overview of other measures is collected in [25] and [26].

Chapter 3

The Data

All analyses contained in this paper work with *television audience monitoring data* measured in the Czech Republic. They were produced by *Mediaresearch a.s.* for *ATO (Association of Television Organizations)* in the years 2002-2006.

3.1 Purpose

Television data are mainly used by two types of subjects to satisfy their needs:

- They provide television networks with feedback about the success of broadcasted programs and overall schedule. It retrospectively influences program structure to address more viewers.
- They enable advertising companies to make plans for campaigns by placing commercials into suitable blocks to strike the maximum number of people from the *target group*. Media representatives of television companies also calculate prices for broadcasting time based on this data.

This means that the television advertising market works thanks to these data. They provide a quantitative measurement of how much are single programs and commercials are watched, and this is the most important factor influencing the incomes of television metworks because they are primarily derived from commercial advertising. The more people watch the commercial, the more money the network gets for its broadcasting. This simple principle forces television networks to broadcast programs that have great mass-popularity yet are sometimes less culturally valuable.

3.2 Production

The preparation of the *peoplemeter* project is a very complex process even in such a small (10 mil.) country like the Czech Republic. It requires a well

balanced selection of participating respondent households to preserve sufficient *representativeness* of such a sample relative to the whole country. And although *technical aspects* should not be overlooked, there is one more crucial segment. It could be called *maintenance* and deals with the prompt identification of extraordinary conditions and persistent motivation of respondents to high-quality cooperation.

3.2.1 Selection of respondents

This part of the project plays an important role in securing representative coverage of the whole population in a chosen sample. In the Czech Republic, the current project guarantees data from at least 1.200 households, which is approximately 3.200 respondents. It corresponds to 0,03%, which is not much at first sight.

The very first phase consists of *address collection*. 250.000 addresses were collected within previously given sample points on the whole area of the country. Among them, we have chosen candidates for the *establishment survey* according to size and location of place they live in. The establishment survey is a very comprehensive questionnaire answered by 12.000 households, finding out their sociodemographical facts like education, age, gender, marital status, work activity and their family constitution (counting number of kids, monthly incomes, social subsumption etc.). They also filled up information about their television behavior and domestic audio-video facilities.

Some of the respondents that answered the establishment survey became candidates for *recruitment*. They were chosen according to panel norms based on the census made by *ČSÚ (Czech Statistical Office)* to represent whole population in attributes that influence watching television. If they agreed, they were invited to cooperate with us. The establishment survey is actualized three times a year by another 2.000 sample acquired by a *continuous survey*, which is the basis for an annual one-quarter renewal of the panel.

3.2.2 Technical aspects

All respondent households have built-in *probes* in their television devices. They monitor the usage of all *TV sets, videos, satellite receivers, DVD players, game consoles etc.* Each stable respondent or temporary guest, who watches the TV registers himself by logging in via remote control. When he finishes watching, he logs off. The probe makes a record of this activity and joins it with the information about the channel watched. It is consecutively transferred to the *communication unit* that collects data from all probes in the household and sends them via *GSM network* to the *collecting center*, which accumulates data from all households for further processing.

Every morning at 6 o'clock, all individual television activities are imported into a central database, connected with information about respondents and their families, broadcasted programs, commercials etc. Then the data is cleaned (EG EXTREMELY LONG-LASTING ACTIVITIES ARE FILTERED OUT), weighted to fit sizes of given population groups (EG BY GENDER, AGE, REGION, EDUCATION) and a few statistical parameters are checked, to make sure it went alright. Then it is prepared for usage.

3.2.3 Maintenance

We have already mentioned the importance of selecting suitable households to fit the global demand on representativeness. It qualifies the measurement as a valid sociological inquiry but still does not guarantee permanent quality of the data, since respondents' morale tends to descend during long-term cooperation. There are two main ways to improve the overall quality of the measurement concerning the behavior of respondents:

- *Active motivation* of the cooperating subject by emphasizing the importance of the project and it's ability to influence the broadcasting schedule.
- *Regular data checks* that would identify any discrepancies in respondents' television behavior indicating either loss of interest or technical problems.

Both these mechanisms have to work hand in hand. They entail daily communication with households, backward verification of measured entries, comparison of contemporary indicators with historical data, and filtering of impossible situations like continual 8-hour long viewing.

3.3 Content

The data sets produced as a result of the above-described mechanisms, securing the highest possible correspondence with reality, consist of several parts:

- *Sociodemographical data* that contain information about individual respondents, including gender, age, education, social class, income, work activity, usage of the internet, declared usage of TV etc., and about their households. It is common for all its members and includes region, size of the city they live in, number of kids living in the family, number of TV sets and videos, etc.
- Records of *individual activity* characterize when single respondents turn on/off the television, when they switched channels and what channel they watched. It is measured with *one-second precision.*
- *Program data* contain the description of individual programs broadcasted by all major stations. It includes the program's title, genre, original language mutation, target group, etc and the measured ratings for chosen population groups.

- *Commercial data* are analogous to program data but they describe broadcasted commercials. They contain information about their advertising company, product type, brand, creative agency etc.
- *Self-promotion data* is the third descriptive set, in this case used by television networks. It holds information about film trailers and excerpts of other future programs broadcasted to attract its viewers.
- *Aggregated minutes* contain a matrix of pre-computed ratings of all 1440 minute daily intervals for chosen target groups.
- *Hydrometeorological data* follow the development of weather and temperature in regions.
- *Life Style Survey (LSS)* is an additional inquiry running continuously during the whole project. It is a thorough questionnaire asking both individual respondents and complete families as represented by their heads about life attitudes and priorities, media behavior, preferences and consumer habits, favorite products and brands, and average consumption of many different goods.

3.4 Common usage

All introduced types of data create a very rich foundation for variety of different analyses and evaluations. Some of them are used in practice very often. They report program ranking, average daily, weekly, monthly and annual viewing, campaign evaluations and efficiency of its cost, channel shares, etc. Most of them could be contained in the category of data retrieval. This means that source data sets are simply read, filtered, accumulated and sorted. In the next chapter, we will set up some more complicated analyses that will be categorized as data mining tasks.

3.4.1 Frequently used indicators

Television audience measurement uses very often several indicators. Let's explain the most important ones:

- *Rating* is the essential and most often used value, which describes the percentage of all examined people that were watching a certain channel (or TV at all) during a particular period (eg time interval, TV program, commercial etc.). In the simplest case of the indivisible interval of, let's say, 1 second, its rating on a given channel is counted as C_w/C_a, where C_w represents the number of examined people that were watching a given channel during the chosen period and C_a is number of all examined people. These counts can be represented by either raw *counts of respondents* or by *weighted projections*.
  If rating refers to a longer period, when people can change their state

towards the analyzed channel (watch for a while and then not) or even more discontinuous intervals, the overall rating is an average of ratings of single 1-second intervals.

- *Share* is closely related to rating but it only takes into account people that were watching TV during some period. Therefore, the sum of the shares of all channels is 1 (100%).
- *Reach (OTS - opportunity to see)/NetReach/Frequencies* are indicators used to evaluate viewing of longer interval or more of them. Reach requires a definition of the reached respondent. It says what viewing condition for an analyzed interval has to be accomplished in order to count the respondent as a hit. This condition can, for instance, be one third of the interval, 5 minutes in all, etc. Frequencies deal with several intervals and group people that were hit by at least n intervals. NetReach is the same as frequency for $n = 1$.
  Reach for 1 interval is counted as C_h/C_a, where C_h is number of people hit by a given interval (according to the definition of reach). Frequency equals C_{h_n}/C_a, where C_{h_n} is the number of people hit by at least n intervals.
- *Affinity* is a quotient R_{TG_1}/R_{TG_2}, where R_{TG_i} is the rating of a target group TG_i. It indicates, which target group was more suitable (accepted better certain program).
- *Index* similarly compares ratings of different channels and is counted as R_{C_1}/R_{C_2}, where R_{C_i} stands for rating of channel C_i
- *CPP (CPRP) - cost per (rating) point* is counted as P/R, where P stands for price of, typically, broadcasting a commercial and R is its rating. CPP is a sort of normalized cost of a commercial or whole campaign.
- *CPT - cost per thousand* is analogous to CPP with the difference that it is related to thousand of people from population of particular target group.

3.4.2 Common analyses

This section aims to demonstrate a few typical analyses of television viewing based on peoplemeter data, that could be included in the category of data retrieval.

Weekly top 10 programs

Probably the most frequent type of analysis related to television viewing is the *top list*. It is the simplest way to present the success of broadcasted programs in various time periods. In fact, top list is a trimmed table of programs sorted by their Rating in descending order. It is used both by television networks for their internal needs and by newspapers and magazines for the public. Figure 3.1 demonstrates such an example.

Rank	Program	Channel	Day	Beginning	End	Rating (%)
1	Tennis USA vs. CHI (OG ATHENS 2004)	ČT2	Sunday	21:38:41	21:51:41	14.2
2	News	ČT1	Sunday	19:14:50	19:32:29	14.0
3	News	ČT1	Friday	19:14:50	19:39:10	12.8
4	News	ČT1	Saturday	19:14:50	19:34:59	12.5
5	News	ČT1	Tuesday	19:14:50	19:38:17	12.2
6	Chalupáři (TV Series)	ČT1	Wednesday	20:03:14	20:42:10	12.0
7	Tennis USA vs. CHI (OG ATHENS 2004)	ČT2	Sunday	20:33:41	20:44:41	11.9
8	News	ČT1	Monday	19:14:50	19:38:18	11.5
9	News	ČT1	Thursday	19:14:50	19:38:27	10.8
10	News	ČT1	Wednesday	19:14:50	19:37:28	10.8

Figure 3.1: Top 10 programs on public service TV channels during the 34^{th} week of 2004, when the Olympic Games in Athens took place

Share of channels

Another typical usage illustrates the comparison of share for particular media (TV channels) on the market. It often referes to long-term periods; see Figure 3.2.

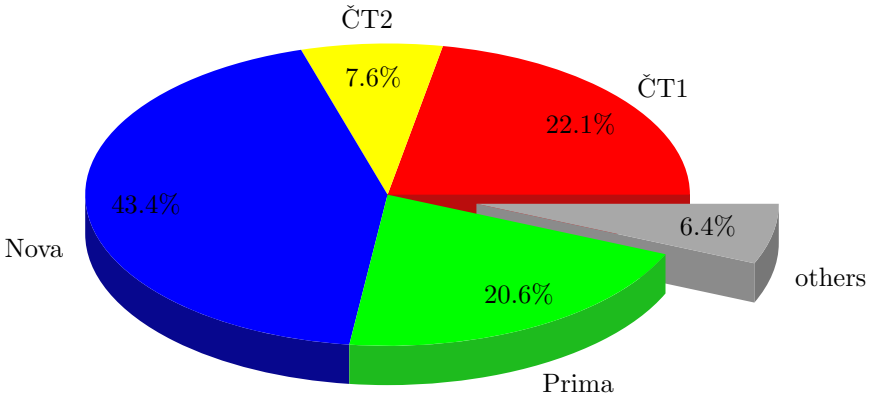

Figure 3.2: Share of channels in year 2003

Campaign evaluation

Advertising agencies buy advertising space in media (IE TIME ON TV, BILLBOARDS OUTDOORS, SPACE FOR BANNERS ON THE INTERNET) to place there advertisement. Media representatives sell this space. Both of these sides need

feedback regarding the effect the emission of advertisement had on the public. It is the basic mechanism that makes the advertising market work. Figure 3.3 and Figure 3.4 demonstrate typical analyses of a TV ad campaign.

Channel	Length	Count	GRP(%)	NetReach(%)	CPP	CPT	Price(Kč)
Nova	30	95	891.9	87.5	18 564	219	16 557 500
Nova	59	1	1.1	1.1	36 215	427	40 700
Nova	60	10	45.3	26.6	36 748	433	1 665 000
Nova		106	938.4	87.9	19 463	229	18 263 200
ČT1	30	53	269.4	68.4	20 430	241	5 503 693
ČT1	60	2	10.9	9.0	38 772	457	420 682
ČT1		55	280.2	69.1	21 140	249	5 924 375
Prima	30	57	343.3	68.5	17 510	206	6 012 000
Prima	60	5	21.2	16.1	33 361	393	708 080
Prima		62	364.6	69.4	18 433	217	6 720 080
Total		223	1 583.2	94.4	19 523	230	30 907 655

Figure 3.3: Summary pivot table (by channel and length of the commercial) referring to a TV ad campaign running in March 2004

Daily course of viewing

Sometimes, it is very important, especially for media themselves, to compare detailed audience curves within a short period of time. It enables them to detect the influences of competing channels on local minima and maxima, as well as the progress of viewing inside a single program itself as shown in Figure 3.5.

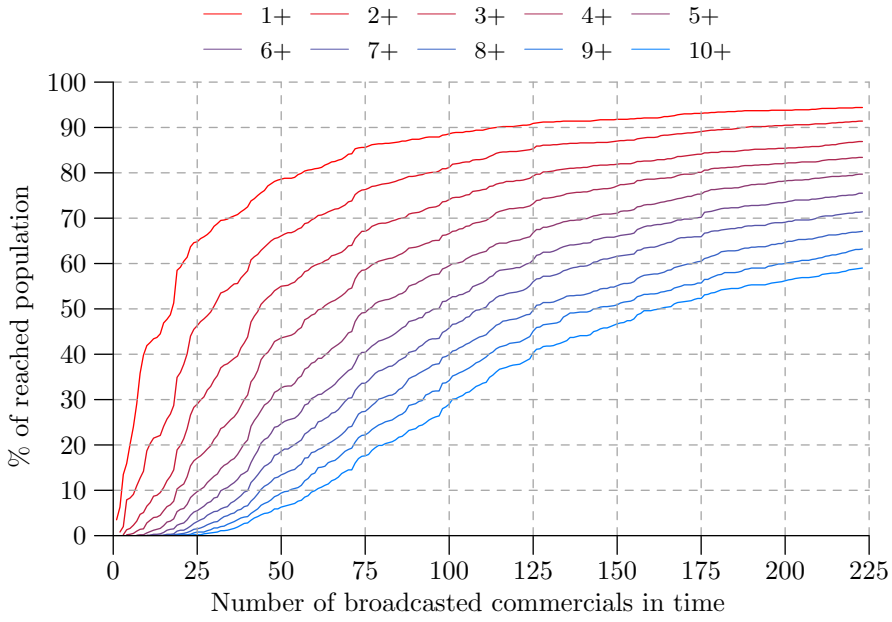

Figure 3.4: This graph demonstrates how single effective frequencies (groups of people reached $n+$ (at least n) times) grew during the campaign broadcasted in March 2004)

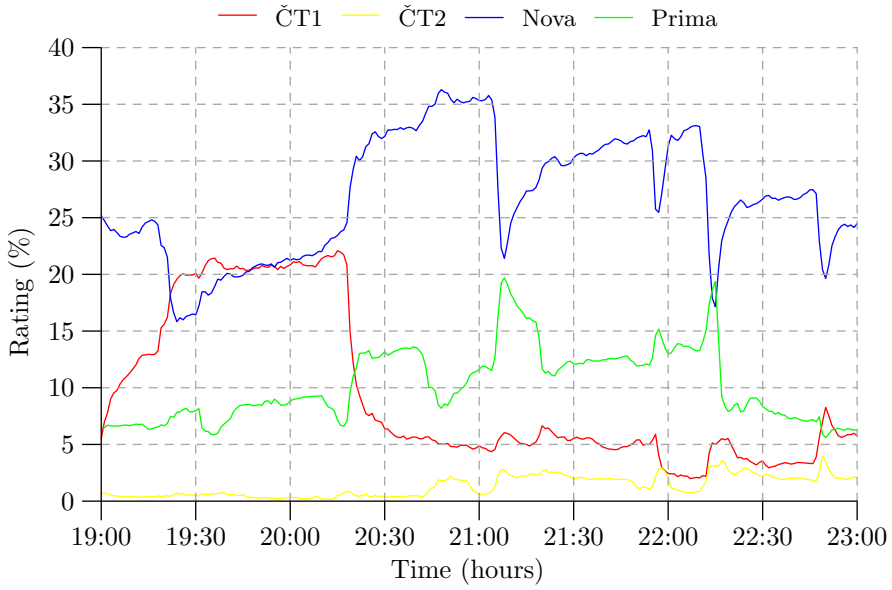

Figure 3.5: Rating progress of main TV channels during *prime time* in Christmas Eve 2003

Chapter 4

Usage

In this chapter, we will utilize our knowledge of methods (section 2.4) and the data itself (chapter 3) to find out what is going on inside of them. We will also try to respect the main guidelines of the CRISP-DM framework (section 2.3). Our aim is to compare the outputs and also the process of exploring the data from the perspective of both business value of discovered relationships and practical usability. The qualities of the used software implementations used will be also mentioned, if needed/important.

4.1 Analyses

Now, we will deal with a sequence of concrete and practically useful data mining tasks. We will use data from the *Lifestyle survey (LSS)* which provide us with huge amount of variables (the questionnaire contains three sections with approx. a thousand questions). The sample of 3.000 respondents also promises interesting hidden coherences. All cornerstones of CRISP-DM used in further paragraphs were described in general in subsection 2.3.1.

4.1.1 CRISP-DM in use

Business understanding Our *goal* is to compare methods similar to GUHA in its *interpretability* and *form of outputs* within the application to television audience data. We want to understand each of them and obtain as much new knowledge about the data as possible. Consecutively, we will compare the results and evaluate their assets to global understanding of the analyzed population and its behavior. The project will be successful if it brings us insight into the Czech TV population. We are interested in its habits, major attributes of main categories; its segmentation in general. The accent is to be put on the comparison of different contexts and situations, in which individual methods should be used due to distinct results.

Data understanding We have described data thoroughly in chapter 3. Our data sources are made of the LSS conducted in autumn 2005. In the first stage, a simple data exploration is made, so that we get to know it better. This step does not extract sophisticated knowledge yet. It helps us to know the quality of individual variables, frequencies of their values, and the measure of missing values.

As GUHA works with discrete variables only, we had to discretize a few continuous variables (eg age) to produce intervals with similar frequencies. In the following step we found out that several variables do not give us much interesting information (sometimes referred to as *entropy* computed as $-\sum_{i=1}^{n} \log_2 Fr_v(i)$ where $Fr_v(i)$ is the frequency of value i in variable v) about the sample, because the distribution of its values approaches 100% for one value and 0% for others. We deleted these variables to simplify the input and also to increase the chance of getting interesting rules. The threshold was experimentally set to 75%, so none of the variables satisfying $max_i(Fr_v(i)) \geq .75$ enters next stages of the process. The marking of the negative values "N/A", "NOT STATED", "NEVER" OR "NO" as a *missing value* has the same purpose, since in the data mining process they can be treated as special values with no relevancy. Finally, we have reduced the data set to 300 variables, which makes the situation significantly simpler.

Data preparation Although this part of the process is very important and has its bottlenecks and interesting technical aspects that may affect the complexity of each data mining task, its description is not the aim of this paper. For further information see section 2.6.

Modeling and Evaluation In this section, we will use software implementations and import prepared data into them. We will run concrete analyses and iterate the process of definition and evaluation (individual procedures will be run and their outputs will be assessed).

Deployment The fruits of our work, specifically our task defined above, lead to chapter 3 and chapter 5. The most important pieces of new knowledge about the data and performance of the algorithms used are discussed there.

4.2 Guha in action

4.2.1 Software implementation

GUHA methods have been implemented several times over its aged history. The last very nice attempt resulted in the application *LISp-Miner* [20] programmed at the *University of Economics, Prague* and it was used for all GUHA analyses. It communicates with data sources via *ODBC* and has these key features:

- automatic generation of categories for continuous variables based on either the sizes of intervals (equidistant style) or the number of objects inside of them (equifrequent style),
- user-defined selection of candidate attributes for antecedents and succedents,
- definition of cedent length (number of atoms employed),
- constraint for output hypotheses made of GUHA quantifiers and the absolute number/percentage of objects satisfying it,
- output export.

LISp-Miner implements *deleting*, *secured*, and *optimistic* strategies of handling missing values. We will preferentially use the most strict - *secured* version to minimize risks, but with a full awareness that we also may lose a few interesting rules. IT, FOR EXAMPLE, HAPPENS IN SITUATIONS WHEN THE MISSING VALUE IS IDENTIFIABLE WITH NEGATION OF THE TESTED PROPERTY. Nevertheless, the size of the data enforces a holistic approach.

4.2.2 First round

The initial intuitive idea says to let the software explore the complete data set. Just to set reasonable thresholds for quantifiers and present computed results. However, the user very soon finds that when handling such complex data as LSS results, the majority of discovered relationships lack interestingness. This is caused mainly by two factors:

- Lots of rules simply describe obvious patterns. We can call it *background knowledge*.
- Lots of rules describe patterns appearing by chance, their validity is just a coincidence, not a causality.

AFTER THE FIRST ITERATION OF THE DATA MINING TASK, WHEN WE WERE INTERESTED IN RULES MADE OF 1-ATOM CEDENTS WITH *support* $\geq 50\%$ AND *founded implication* HAVING $p \geq .75$ WE OBTAINED ABOUT 150 RULES. IN SPITE OF THAT, MOST OF THEM LOOK LIKE THE FOLLOWING:
- 100% OF PEOPLE, WHO DRANK LAGER DURING LAST YEAR, DRANK BEER. OR
- 82% OF PEOPLE, WHO USUALLY SEND AND RECEIVE SMS, HAVE EATEN ICE-CREAM.

The first type is only a result of logical coherences inside the survey, because some questions are tied to others. The second one looks more interesting but we have to handle it carefully. Founded implication is a quantifier that helps to trace what properties those people satisfying an antecedent have besides that. Nevertheles, categories with dominant representation in the population actually occur together. If we want to discover something new, we have to filter out such moments.

4.2.3 Second round

Fortunately, we have other quantifiers, which might help us to accomplish such a task. Before using them, we will also get rid of maximum logical dependencies arising from the construction of the questionnaire. Thanks to that, we will not receive the first type of obvious patterns. On the other hand, something may be lost. Since we do not want to mechanically read hundreds of rarely interesting statements, we choose this way.

Now, we can advance to a deeper dissection of causality. The *above average* quantifier seems to be a good candidate for the job. It compares the confidence of the rule and the relative frequency of a succedent (or *lift*, as is often used in literature) and therefore it testifies how features from an antecedent affect the appearance of features conforming to a succedent (and vice versa, because it is symmetric). In the fact, it compares factual state and hypothetical estimation, if cedents were statistically independent. GUHA also defines a quantifier based on *Fisher's exact test*, which examines the dependency of two discrete variables, but a huge amount of rules pass it even on very low (less than 10^{-5}) levels.

ALTHOUGH AGE IS ONE OF MOST ESSENTIAL SOCIODEMOGRAPHICAL ATTRIBUTES, WE DELETED ITS VARIABLE FROM OUR MATRIX, SINCE OUR RESULTS IN THE SECOND STAGE WERE FULL OF COMBINATIONS WITH YOUNG PEOPLE (FIRST CATEGORY: 15-25) AND THEIR COMMON CHARACTERISTICS LIKE ELEMENTARY EDUCATION OR BEING SINGLE, AND ELDERLY PEOPLE (LAST 2 CATEGORIES: 65-75 AND 75-85), WHO ARE USUALLY RETIRED OR ECONOMICALLY INACTIVE FOR A LONG PERIOD OF TIME (LAST CATEGORY: OVER 10 YEARS). THE SAME SITUATION HAPPENED WITH SEVERAL OTHERS LIKE MARITAL STATUS AND SOCIAL STATUS (WHICH INCLUDES A CATEGORY OF STUDENTS/APPRENTICES). EXCEPT FOR THESE MORE INFORMATIVE, WE PURGED VERY WEAK (LOW RESPONSE RATE) VARIABLES DESCRIBING THE CONSUMPTION OF REGIONALLY BASED MEDIA THAT STRONGLY CORRELATE WITH REGIONS.

THE USAGE OF THE ABOVE AVERAGE QUANTIFIER GIVES US THE POSSIBILITY OF FASTER INSIGHT INTO THE DATA. SORTING THE RULES WITH HIGH CONFIDENCE SHOWS THE MORE INTERESTING ONES NEAR THE TOP:
- 90% OF PEOPLE, WHO DRINK LEMONADES, EAT ICE-CREAMS, WHICH IS 21% MORE THEN IN THE WHOLE POPULATION (A SIMILAR CORRELATION WAS FOUND WITH EATING SWEETS).
- 80% OF PEOPLE, WHO DRINK WINE, ALSO DRINK BEER. AND IT IS ALSO 21% ABOVE AVERAGE.

NOW, THANKS TO THE NEW FILTER, WE CAN DECREASE THE LOWER BOUND OF SAMPLE SIZE AND READ OTHER INTERESTING STATEMENTS SUCH AS:
- 87% OF PENSIONERS NEVER PLAY COMPUTER GAMES AND IT IS 1.6 TIMES MORE OFTEN THAN AMONG THE WHOLE POPULATION.
- 92% OF PEOPLE, WHO DO NOT INCLUDE SELF-EDUCATION AMONG THEIR LEISURE ACTIVITIES, ALSO DO NOT USE THE INTERNET, WHICH IS 36% MORE THAN IN TOTAL.

4.2.4 Final round

With the improved methodology of using GUHA procedures together with better understanding of the data and their "built-in" mechanisms (not forgetting the fact that it describes the population we know from inside), we can use more computationally complex procedures allowing 2-atom antecedents. FIGURE 4.1 CONDUCTS MOST INTERESTING OF THEM.

4.3 Association rules in action

4.3.1 Software implementation

We will use *Weka* [28] to extract *association rules*. It is also the product of an academic endeavor, this time coming from New Zealand's *University of Waikato*. It implements plenty of data mining methods and is written in *Java*. Contrary to *LISp-Miner*, *Weka* implements a *non-deleting* approach to missing values only. As we have used the *secured* version previously, we may, thanks to it, obtain some rules that would not fit it.

4.3.2 Metrics

Apart from that, *Weka* comes with important extensions of the generic generation of association rules. A user may specify the *upper* and *lower* bound for support and a *step*. The algorithm iterates between them going down and tries to find a demanded *number of rules*. Then it stops. *Weka* also implements three more metrics (quantifiers) in addition to *confidence* in the Figure 4.2. *Lift* is, in fact, equivalent to the *above average* quantifier. It compares the actual frequency of both φ and ψ occurring together with the hypothetical probability of it, if they were independent. While *lift* is computed as their ratio, *leverage* is the difference. So, these two metrics are very closely related, they differ in the scale only. And finally *conviction*, similarly to *lift*, compares hypothetical and actual frequency of cases satisfying φ and $\neg\psi$. According to results from [6] and [4], all of the other metrics are monotone in confidence (so confidence is included in them and can be recast) and whereas *lift* reports statistically motivated co-occurrence (as it is symmetrical), *conviction* is directional and identifies factual implication.

4.3.3 Application

For mining association rules, we used previously conducted data transformations and experience acquired when using GUHA - the very last data set of independent variables will be employed. We let *Weka* generate the first 100 rules for all metrics via the iterative decreasing of its lower bound. And what did we get?

Antecedent	Succedent	Conf	AA	Supp
Married male	Greatest interest in news (as TV program)	83%	26%	708
Housewife	Daily housework	80%	50%	1116
Woman	Daily housework	74%	37%	1177
Apprenticed	Never using the internet	76%	30%	894
Retired woman	Never actively engaged in sports	75%	87%	356
Woman, typology: "searching"	Eating cereal products	77%	89%	102
University educated, high living standard	Using payment cards	78%	103%	63
Male of age 75-85	Most often consumed beer packing: bottled	75%	105%	52
Cultural typology: strongest, 51-100 employees in the company	Using payment card	75%	97%	55
Married	Paying extra insurance	84%	11%	1432
Male	Drinking beer	81%	23%	1208
Married w/ access to the internet	Attending internet chats	55%	135%	339
Married w/ access to the internet	Attending internet chats	55%	135%	339
Subjective assessment of health: very good	Single	52%	108%	433
Daily chatting via the internet	15-25 years old	80%	369%	32
Attending disco/dance party at least once a week	15-25 years old	76%	345%	126
Daily self-education	Access to the internet	77%	102%	250
Proprietor w/o employees	Social profile: C	82%	205%	122
EGP category: housewife	Parental leave	75%	1790%	77
Social profile: A	Cultural life typology: strongest	68%	133%	119
Social profile: A, 35-45 years old	University education	81%	741%	29
Elementary education (incl. unfinished), living alone	Social profile: E	84%	119%	51
Interested in philosophical questions, social profile: A	Town city: 100.000+	79%	289%	23
Searching excitement in TV, both parents, kids and other relatives living together	Social profile: E	78%	104%	21
Not "believing in miracles", single parent w/ kids	Apprenticed	77%	102%	20
Elementary education (incl. unfinished), both parents w/ kids	Single	77%	182%	247

Figure 4.1: Strong and interesting relationships found by Guha

Name	Condition
confidence	$p \leq \frac{a}{a+b} = \frac{Fr(\varphi\&\psi)}{Fr(\varphi)}$
lift	$p \leq \frac{a(a+b+c+d)}{(a+b)(a+c)} = \frac{Fr(\varphi\&\psi)}{Fr(\varphi)Fr(\psi)}$
leverage	$p \leq \frac{a}{a+b+c+d} - \frac{(a+b)(a+c)}{(a+b+c+d)^2} = Fr(\varphi\&\psi) - Fr(\varphi)Fr(\psi)$
conviction	$p \leq \frac{(a+b)(b+d)}{b(a+b+c+d)} = \frac{Fr(\varphi)Fr(\neg\psi)}{Fr(\varphi\&\neg\psi)}$

Figure 4.2: Metrics in *Weka* for $\varphi \Rightarrow \psi$

The very first (with greatest support) computed rule ρ with $conf(\rho) = 1$ is:
- People, who use aftershave, are male.
At first sight, it was a surprise that we did not get this rule with Guha. But the explanation is very simple. Although 1070 males declared they use aftershave, 1600 female did not answer the question, which is represented by a missing value in the data. As we used a different (more careful) method of handling *m.v.* in Guha, the confidence resulted in 40%, which was below the threshold.

Because each implementation is equipped with different missing values conversion strategies, we cannot fully compare them. Nevertheless the differences caused by this fact are not diametrical. After the first run, we have decreased the lower bound of support and demanded higher metrics. Especially *conviction* brought us new observations, as its alternative is not included in *LISp-Miner*. The following table Figure 4.3 summarizes the most interesting rules found by *Weka* (and not found by *LISp-Miner*) after several iterations of parameters definition.

Antecedent	Succedent	Metric/value	Supp
Spending less than 400 crowns/month for mobile communication	Uses pre-paid mobile cards	conf=84% lift=146%	1459
Never surfing the web w/o concrete goal	Never playing web games/competitions	conf=72% lift=375%	398
Never sending SMS via the internet	Never playing web games/competitions	conf=66% lift=342%	373
Does not feel healthy	Statement "Nearly permanently feeling tired, neither weekends help"	conf=44% lift=1619%	49
Surfing the web w/o concrete goal every day	Seeking information for personal usage every day	conf=62% lift=1142%	38
Downloading ring tones	Downloading Java games	conf=45% lift=961%	91
Seeking information for professional usage (web) every day	Seeking information for personal usage every day	conf=42% lift=772%	75
Taking care of livestock	Taking care of pets	conf=55% lift=617%	55
Woman	Housewife	conf=76% conv=227%	1215
Both parents w/ kids	Using SMS technology in GSM network	conf=81% lift=160%	1280
Visiting galleries/exhibitions (once+ per month)	Visiting muzeum/zoo/castles (once+ per month)	conf=43% conv=145%	43
Not respecting family tradition for celebrations	Disagree w/ "beloved partner should be married"	conf=35% conv=140%	34
Visiting theater/ballet (once+ per month)	Visiting cinema (once+ per month)	conf=32%, lift=133%	50
Drinking cola 4-6 times a week	Visiting cinema (once+ per month)	conf=30% conv=129%	36
Most favorite distilled beverage: whiskey/bourbon	Spending more than 800 crowns for mobile communication	conf=28% conv=128%	37
"I am not aware of my limitations often."	"If things go other way than I want, I become angry."	conf=29% conv=128%	68

Figure 4.3: New relationships found by association rules (based on confidence, lift and conviction)

Chapter 5

Conclusion

We have tried to map the basic assumptions and typical purposes of data mining tasks together with a definition of a standardized process for treating them. Understanding two similar methods and a practical example of their usage might form a good starting point for further comparisons, especially concentrated on algorithms that produce rules.

5.1 The methods and implementations

We have gone through a data mining process with a specific data set from a lifestyle survey carried out upon respondents of television audience measurement. It was conducted in two academic applications, using implementation of two similar methods with very different histories. While association rules were primarily designed for simple market basket data analysis with the accent on effectiveness, GUHA stands on a very solid theoretical foundation and originates in a much distant past.

Nevertheless, huge tests, applications and efforts of many people have enriched the simple initial principle of association rules, which made it usable in more complicated situations. Thus, it approached GUHA, which, on the other hand, lacks of massive usage. It is also hard to separate the method and concrete implementation, as many added features come from isolated data mining areas. Put together, experiences gained during the data processing and their comparison highlight several recommendations as to how the best of each method could be combined:

- Association rules in *Weka* are computed much faster (with a better implementation of the algorithm) that allowed the computation of 3-atoms-long antecedents, while *LISp-Miner* was not able to compute such rules for the data of tested size within two days, which makes it, in this situation, unusable.

- *LISp-Miner* employs 3 strategies to convert 3x3 contingency tables with missing values, *Weka* uses only one, which is different to them. All of them might be useful in diverse situations depending on the origins of missing values.
- *LISp-Miner* implements a much larger number of quantifiers and allows the combination of their constraints for resulting rules. Especially a statistically motivated quantifier based on *Fisher's test* has its solid theoretical basis. However, our data did not confirm its relevance, *above average* did a better job. *Weka* also comes with an interesting extension: *conviction metric*.
- *LISp-Miner* is very strong in the definition of generated hypotheses. The user can determine the complexity of cedents (their length in atoms) and also the manner of atom generation (size of subsets/cuts of allowed values).
- The support boundary iterator together with a specification of number of requested rules implemented in *Weka* is very practical, particularly while executing first runs.

5.2 The data and results

Because every data mining task should report the results and a significant part of the work in this document is constituted of practical application, we will emphasize a few nuggets found inside the data. Others are in the tables in chapter chapter 4.

- The internet is a medium used jointly for both professional and personal purposes.
- Internet users can be divided into two groups: practical users and people who use it for entertainment.
- Cultural activities tend to occur together.
- Higher education influences the interest in news, culture and the internet.
- Favorite beverages might serve as indicators of other habits.

5.3 Proposals for further investigation

In this document, we have tried to connect two worlds: the academic field, from which interesting automatized knowledge discovery methods have arisen, and the world of applied research with its practical demands. This endeavor was put into a framework, which introduces the basic principles and categorization of data mining as a fast growing branch incorporating many fields of mathematics and computer science. But, this is just the beginning.

We have concentrated on one segment of peoplemeter data and two methods only. But usual work, even with this type of data, includes other data sets, especially those concerning continuous audience variables. Besides that, there are several other methods producing (or capable of producing) a similar type of knowledge as association rules and GUHA, which should be also compared, but are beyond the scope of this introductory text. In this context, we should mention at least: algorithms associations-like algorithm `Tertius`, *inductive logic programming* (ie `Golem`, `Foil`), *genetic algorithms*, and not forgetting methods for *discretization*. Some other approaches, ie those building *decision trees*, are not primarily designed to extract rules, but might be extended to do so.

I hope, that this thesis inspires the people of praxis to use the methods and tools described in it, and provides feedback about their usability to those people, who invent and implement them.

References

[1] Rakesh Agrawal, Tomasz Imielinski, and Arun Swami. Database mining: A performance perspective. In Nick Cercone and Mas Tsuchiya, editors, *Special Issue on Learning and Discovery in Knowledge-Based Databases*, volume 5, pages 914–925. Institute of Electrical and Electronics Engineers, Washington, U.S.A., 1993.

[2] Rakesh Agrawal, Tomasz Imielinski, and Arun N. Swami. Mining association rules between sets of items in large databases. In Peter Buneman and Sushil Jajodia, editors, *Proceedings of the 1993 ACM SIGMOD International Conference on Management of Data, Washington, D.C., May 26-28, 1993*, pages 207–216. ACM Press, 1993.

[3] Rakesh Agrawal and Ramakrishnan Srikant. Fast algorithms for mining association rules. In Jorge B. Bocca, Matthias Jarke, and Carlo Zaniolo, editors, *Proc. 20th Int. Conf. Very Large Data Bases, VLDB*, pages 487–499. Morgan Kaufmann, 12 1994.

[4] R. Bayardo, R. Agrawal, and D. Gunopulos. Constraint-based rule mining in large, dense databases, 1999.

[5] Sergey Brin, Rajeev Motwani, and Craig Silverstein. Beyond market baskets: Generalizing association rules to correlations. In Peckham [23], pages 265–276.

[6] Sergey Brin, Rajeev Motwani, Jeffrey D. Ullman, and Shalom Tsur. Dynamic itemset counting and implication rules for market basket data. In Joan Peckham, editor, *SIGMOD 1997, Proceedings ACM SIGMOD International Conference on Management of Data, May 13-15, 1997, Tucson, Arizona, USA*, pages 255–264. ACM Press, 05 1997.

[7] Pete Chapman, Julian Clinton, Randy Kerber, Thomas Khabaza, Thomas Reinartz, Colin Shearer, and Rüdiger Wirth. *CRISP-DM 1.0, Step-by-step data mining guide*, 2000. More information available at `http://www.crisp-dm.org/`.

[8] Usama M. Fayyad. Mining databases: Towards algorithms for knowledge discovery. *IEEE Data Engineering Bulletin*, 21(1):39–48, 1998.

[9] Petr Hájek. The GUHA method and mining associational rules. In *CIMA*, 2001.

[10] Petr Hájek. On generalized quantifiers, finite sets and data mining. In Klopotek et al. [19], pages 489–496.

[11] Petr Hájek. Metoda GUHA v minulém století a dnes. In Václav Snášel, editor, *Znalosti 2004*, pages 10–20. FEI, VŠB, Ostrava, 2004.

[12] Petr Hájek, Ivan Havel, and Metoděj Chytil. GUHA - metoda systematického vyhledávání hypotéz. *Kybernetika*, 2(1):31–47, 1966.

[13] Petr Hájek, Ivan Havel, and Metoděj Chytil. The GUHA method of automatic hypotheses determination. *Computing*, 1(4):293–308, 1966.

[14] Petr Hájek, Ivan Havel, and Metoděj Chytil. GUHA - metoda systematického vyhledávání hypotéz ii. *Kybernetika*, 3(5):430–437, 1967.

[15] Petr Hájek, Ivan Havel, and Metoděj Chytil. Problém obecného pojetí metody GUHA. *Kybernetika*, 4(6):505–515, 1968.

[16] Petr Hájek and Tomáš Havránek. *Mechanizing Hypothesis Formation (Mathematical Foundations for a General Theory)*. Springer-Verlag, New York, 1978. Internet edition of the monograph at http://www.cs.cas.cz/~hajek/guhabook.

[17] David Hand, Heikki Mannila, and Padhraic Smyth. *Principles of Data Mining*. The MIT Press, Cambridge, Massachusetts, 2001.

[18] Jochen Hipp, Ulrich Güntzer, and Gholamreza Nakhaeizadeh. Algorithms for association rule mining – a general survey and comparison. *SIGKDD Explorations*, 2(1):58–64, July 2000.

[19] Mieczyslaw A. Klopotek, Slawomir T. Wierzchon, and Krzysztof Trojanowski, editors. *Intelligent Information Processing and Web Mining, Proceedings of the International IIS: IIPWM'03 Conference held in Zakopane, Poland, June 2-5, 2003*, Advances in Soft Computing. Springer, 2003.

[20] Homepage of lisp-miner project at http://lispminer.vse.cz.

[21] J. Mata, J. L. Alvarez, and J. C. Riquelme. An evolutionary algorithm to discover numeric association rules. In *SAC '02: Proceedings of the 2002 ACM symposium on Applied computing*, pages 590–594, New York, NY, USA, 2002. ACM Press.

[22] Tom M. Mitchell. *Machine Learning*. McGraw-Hill, Singapore, 1997.

[23] Joan Peckham, editor. *SIGMOD 1997, Proceedings ACM SIGMOD International Conference on Management of Data, May 13-15, 1997, Tucson, Arizona, USA*. ACM Press, 1997.

[24] Craig Silverstein, Sergey Brin, and Rajeev Motwani. Beyond market baskets: Generalizing association rules to dependence rules. *Data Mining and Knowledge Discovery*, 2(1):39–68, 1998.

[25] P. Tan and V. Kumar. Interestingness measures for association patterns: A perspective, 2000.

[26] P. Tan, V. Kumar, and J. Srivastava. Selecting the right interestingness measure for association patterns. In *Proceedings of the Eight ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2002.

[27] Geoffrey I Webb and Songmao Zhang. Beyond association rules: Generalized rule discovery, 2003. submitted for publication, see: http://www.csse.monash.edu.au/ webb/.

[28] Homepage of weka project at http://www.cs.waikato.ac.nz/ml/weka.

[29] Ian H. Witten and Eibe Frank. *Data Mining, Practical Machine Learning Tools and Techniques with Java Implementations*. Morgan Kaufmann Publihers, San Diego, California, 2000.

[30] Kim Won. Data mining is not against civil liberties. Letter, June 2003. More information available at http://www.acm.org/sigkdd/.

Index