Algorithms for knowledge discovery using relation identification methods

Jakub Mikołaj Tomczak

School of Engineering
Blekinge Institute of Technology
Box 520
SE – 372 25 Ronneby
Sweden
This thesis is submitted to the School of Engineering at Blekinge Institute of Technology in partial fulfillment of the requirements for the degree of Master of Science in Computer Science. The thesis is equivalent to 24 weeks of full time studies.

Contact Information:
Author:
Jakub Mikołaj Tomczak
E-mail: jakubmkt@gmail.com

University advisors:
Jerzy Świątek, Professor
Dept. of Computer Science and Management
Wrocław University of Technology, Poland

Ludwik Kuźniarz, Assistant professor
Dept. of Systems and Software Engineering
Blekinge Institute of Technology, Sweden

School of Engineering
Blekinge Institute of Technology
Box 520
SE – 372 25 Ronneby
Sweden

Internet : www.bth.se/tek
Phone : +46 457 38 50 00
Fax : + 46 457 271 25
Abstract

In this work a coherent survey of problems connected with relational knowledge representation and methods for achieving relational knowledge representation were presented. Proposed approach was shown on three applications: economic case, biomedical case and benchmark dataset. All crucial definitions were formulated and three main methods for relation identification problem were shown. Moreover, for specific relational models and observations' types different identification methods were presented.

Keywords: knowledge extraction from data, relational knowledge representation, relational system, relation identification method, modified coverage knowledge discovery method
For my great uncle, teacher and professor—
R.I.P. Zdzisław Bubnicki.
I would like to thank my supervisor, prof. Jerzy Świątek, for ideas and critical remarks about this thesis. Furthermore, I want to thank my brother and parents for supporting me all the time. Moreover, I am grateful to all my colleges, especially Maciej Zięba, my friend and opponent of this work, for long discussions about problems in field of computer science. Last but not least special regards should be given to the person who has special place in my heart – Natalia.
## Contents

<table>
<thead>
<tr>
<th>Section</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>Abstract</td>
<td>iii</td>
</tr>
<tr>
<td>Dedication</td>
<td>iv</td>
</tr>
<tr>
<td>Acknowledges</td>
<td>v</td>
</tr>
<tr>
<td>Table of contents</td>
<td>vi</td>
</tr>
<tr>
<td>Symbols and abbreviations</td>
<td>x</td>
</tr>
<tr>
<td><strong>1 Introduction</strong></td>
<td>1</td>
</tr>
<tr>
<td>1.1 Problem placement</td>
<td>1</td>
</tr>
<tr>
<td>1.2 Literature review</td>
<td>1</td>
</tr>
<tr>
<td>1.3 Problem background</td>
<td>2</td>
</tr>
<tr>
<td>1.3.1 Data, Information, Knowledge</td>
<td>2</td>
</tr>
<tr>
<td>1.3.2 Knowledge System, Knowledge Object, Model</td>
<td>4</td>
</tr>
<tr>
<td>1.3.3 Reasoning paradigms. Inductive learning</td>
<td>4</td>
</tr>
<tr>
<td>1.3.4 Identification</td>
<td>5</td>
</tr>
<tr>
<td>1.3.5 Main goal of the work</td>
<td>6</td>
</tr>
<tr>
<td>1.4 Research questions and outcomes</td>
<td>7</td>
</tr>
<tr>
<td>1.4.1 Research questions</td>
<td>7</td>
</tr>
<tr>
<td>1.4.2 Outcomes</td>
<td>7</td>
</tr>
<tr>
<td>1.5 Work contents</td>
<td>8</td>
</tr>
<tr>
<td><strong>2 Basic definitions and problem statement</strong></td>
<td>10</td>
</tr>
<tr>
<td>2.1 Relational object</td>
<td>10</td>
</tr>
<tr>
<td>2.2 Partial relation</td>
<td>12</td>
</tr>
<tr>
<td>2.3 Relational model</td>
<td>13</td>
</tr>
<tr>
<td>2.4 Dataset description</td>
<td>14</td>
</tr>
</tbody>
</table>
## Applications of relation identification methods

### 7.1 Relational knowledge representation in economy
- **7.1.1 Problem description**
- **7.1.2 Relational model**
- **7.1.3 Conclusions**

### 7.2 Relational knowledge representation in biomedical case
- **7.2.1 Problem description**
- **7.2.2 Relational model**
- **7.2.3 Results**
- **7.2.4 Comparison**
- **7.2.5 Conclusions**

### 7.3 Relational knowledge representation in benchmark dataset
- **7.3.1 Problem description**
- **7.3.2 Relational model**
- **7.3.3 Results**
- **7.3.4 Comparison**
- **7.3.5 Conclusions**

## Conclusions

### 8.1 Final remarks

### 8.2 Future work

## Bibliography
Symbols and abbreviations

\[ x \in X \subseteq \mathbb{R}^S \] – input vector
\[ y \in Y \subseteq \mathbb{R}^L \] – output vector
\[ R(x, y) \] – relation
\[ w[\cdot] \] – logic value
\[ \Phi(x, y) \] – complex predicate
\[ \phi_m(x, y) \] – simple predicate
\[ \tilde{R}_m(x, y) \] – partial relation (association)
\[ a \in A \subseteq \mathbb{R}^K \] – unknown parameters
\[ R_a(x, y) \] – relational model
\[ F(x, y; a) \] – proposed description
\[ \tilde{F}_m(x, y; a) \] – proposed partial description
\[ R_N \] – dataset (data)
\[ R_{N,1} \] – positive examples
\[ R_{N,0} \] – negative examples
\[ B \] – background knowledge
\[ L \] – language bias
\[ \bar{q} \] – dissimilarity measure
\[ Q(\cdot) \] – criterion
\[ \text{coverage}(\cdot, \cdot) \] – coverage measure
\[ \text{accuracy}(\cdot, \cdot) \] – accuracy measure
\[ \alpha_x, \alpha_w, \alpha_y \] – input, additional and output simple formula
\[ D_a(N) \subset A \] – subset of unknown parameters for which relation covers all positive examples
\[ \Delta_a(N) \subset D_a(N) \] – boundary of \( D_a(N) \)
\[ \Psi(\cdot) \] – an identification algorithm
\[ B \] – body of an implication
\[ C \] – conclusions of an implication
\(d(\cdot, \cdot)\) – metric
\(\text{card}\{\cdot\}, |\cdot|\) – cardinality of a set
\(M(\cdot)\) – mean value
\(\hat{x}\) – soft variable

\(KD\) – Knowledge Discovery
\(DM\) – Data Mining
\(KDP\) – Knowledge Discovery Process
\(RKR\) – Relational Knowledge Representation
\(LAM\) – Logic-Algebraic Method
\(SISO\) – Single-Input Single-Output system
\(MISO\) – Multi-Input Single-Output system
Chapter 1

Introduction

1.1 Problem placement

Nowadays in each minute many data are collected by institutions like hospitals, research laboratories, shops, banks, stores, and so one [20, 39]. In each situation, beside handling and transferring data [47], there is a need of extracting useful information, hidden relationships or knowledge in data, e.g. for creating a credit profile of clients’ groups in a bank, looking for not obvious patterns in chemical or biological substances, classification of illnesses by using different symptoms, seeking unnatural patterns in network’s activity [67]. And because of huge amount of data this need cannot be fulfilled even by experts which are not able to analyze those data in acceptable time and with acceptable quality (so called knowledge acquisition bottleneck [30, 32]).

Therefore there is a big demand for efficient analysis and understanding for huge amount of data. Knowledge discovery (KD) and data mining (DM) methods meet this claim. Moreover, according to this need, KD became one of the most important issues in modern computer science [14, 20, 21, 29, 39, 65].

1.2 Literature review

Knowledge discovery and data mining is a field of computer science which became one of the most important issues in research and industry studies [14, 18, 20, 21, 29, 31, 39, 65]. There are several ways of representing knowledge about systems and object, such as graph-like representations (e.g. decision trees [15, 17, 29, 34, 48, 65], semantic webs [25], ontologies [21]), rules (e.g. clauses,
if-then rules, first-order logic formulas \([15, 14, 29, 30, 34, 48, 65]\), frames \([25]\), uncertainty descriptions \([9, 10, 14, 16, 25, 40, 41, 59]\) (e.g. probability theory, fuzzy sets theory, uncertain variables theory, rough set theory) and other (e.g. neural networks \([48, 65]\), genetic approach \([48]\)). Using one of mentioned representations depends on the situation and needs, however rules and graph-like representations are the most popular and well-established among researchers and industry workers \([20]\).

However, there is also another representation which is called a relational knowledge representation \([4] - [12], [18, 26, 30, 42, 48, 65], [50] - [61]\). There are several approaches to formulate it but all of them are considered about retrieving some unknown knowledge from data. This unknown knowledge is described by formulas of first-order logic (clauses \([15, 18, 30]\), predicates \([4] - [12], [50] - [61]\)) but there are different approaches to discover it. Retrieving clauses are based on inductive concept learning \([15, 18, 30, 34, 48]\), and extracting predicates – on identification methods \([4] - [8], [13], [50] - [61]\). In this work we are concentrated on identification methods.

Nevertheless, identification methods were considered only for very simple relations with numeric attributes. In some papers simple algorithms for static object were presented \([4, 5, 7, 13, 50]\), another for dynamic object \([8]\), and other for static object with additional probabilistic description \([42], [51] - [57], [58]\), where a well-known algorithms like Bayes’ method and Maximum Likelihood method were used. In next works two numerical algorithms were proposed \([60, 61]\).

However, there is a need to create general method for extracting knowledge about the object which can be described by more complex predicates in form of rules. In this work such an algorithm is presented. This approach includes also a relation identification method which is designed for creating a description of numeric part of the relation (description of numeric attributes).

### 1.3 Problem background

#### 1.3.1 Data, Information, Knowledge

Before any considerations some crucial expressions as data, information, and knowledge have to be clearly formulated because in different fields of science they are differently understood.

Data is a collection of examples of some concept which has to be learned \([65]\).
This examples, called also *instances* (e.g. [34, 65]) or *observations* (e.g. [3]), are compound of values of attributes (features) which describe the observed concept. There are three possible values' types of an attribute: *nominal* (names of labels), *numeric* (measured numbers), and *boolean* (true or false value). In other words, data is a collection of unbounded observations [25]. Data can be noised or some values in several instances can be missed. The dataset which contains not noised data without any missing values is called a *learning set*.

*Information*, in the language of *knowledge engineering*, is data which was ordered in some way [25]. For example "temperature of a human is between 0 and 45 Celsius degrees".

**Knowledge** is formed when some conclusions from data and information are drawn. In other words, knowledge could be defined as information with a context [25]. The knowledge divides into three categories:

- *declarative knowledge* – facts, statements and descriptions (e.g. "a patient with temperature higher than 37 Celsius degrees is ill");

- *procedural knowledge* – procedures of reasoning/learning and definitions of predicates, rules (e.g. "if a patient has more than 37 Celsius degrees, then give a medicine");

- *meta knowledge* – 'knowledge about knowledge' (e.g. "if rules and decision tree are available, then use rules").

Besides, there are some other very useful definitions, like *fact*, *background knowledge* and *bias*.

**Fact** could be an observation (instance) which is true without any doubts [25, 30]. It means that there is no noise or any missing values in the observation. However, a fact can be also a statement, which describes a part of a reality, written in a chosen representation form [25, 30]. Facts are derived from fully credible sources like scientific theory or expert.

**Background knowledge** is a set of given knowledge connected with the the considered concept and could be represented in following ways [30, 34, 48]:

- *extensional* – a list of facts about the concept;

- *intensional* – a list of definitions and/or descriptions of a language describing the concept (e.g. description of a rule by other rules, );
• other (e.g. which attributes should be treated as an input, and which as an output).

**Bias** is a mechanism in a learning system to constrain the search of the hypothesis space \[30, 34, 65\]. In this work only language bias is considered which is connected with determining the hypothesis space itself.

### 1.3.2 Knowledge System, Knowledge Object, Model

For avoiding further possible ambiguities, it is useful to exactly describe following expressions: a system, an object, and a model.

A **system** is an expression which is taken into consideration intuitively. In [6] can be found following definition: “The system is some integrated whole in which co-operating, separate components of the system can be distinguished. Functions of components and relations among them create an output functionality of the system. Connections among components define the structure of the system.”

This definition maybe do not describe all possible cases. Nevertheless, it gives a very good idea about the system. In this work by the **knowledge system** we understand a complex concept about some physical quantity in which an input and an output can be distinguished and which are connected by some relation. The system can be considered in many aspects, e.g. an economical aspect [6], a biological aspect [46], a technical aspect [3].

A **knowledge object** will be understood as a part of the knowledge system (component of the system) or an example of the system.

And finally, a **model** can be considered in many ways (e.g. [21]), but in this work it is taken into consideration only in the formalized sense. It means that the model is such description of the system (object) which is understandable for both a human and a machine. Moreover, the model tries to represent a concept of the system as good as possible. The model can be given explicite by the expert or can be a result of some analysis made by an artificial system, e.g. an expert system [5 10 25].

### 1.3.3 Reasoning paradigms. Inductive learning

There are two main paradigms of reasoning [2]:

1. Deduction.
2. Reduction.
A complex considerations about both paradigms can be found e.g. in [2, 48], and here only a short description is showed. In the first approach conclusions about the considered example of the concept are drawn from known theory which describes the example. In the second approach – conclusions about the concept are drawn from a known dataset.

The main method in the reduction reasoning is an *induction*. The induction is a reasoning method which generalizes observations into the output concept.

In modern science the induction is considered as less scientific [2, 43] than the deduction. However, there is no proof yet that the induction should not be treated as a scientific method, especially that this method is widely used in many fields of science. Moreover, the deduction reasoning is very hard to apply in the real-life situations.

On the other hand, instead of reasoning we can talk about *learning*. Learning is a mechanism of acquiring new knowledge from the environment [30]. An *inductive concept learning* could be defined as follows:

*Having a background knowledge, dataset (examples of the concept), and a language which describes a concept (chosen knowledge representation), find a hypothesis (output concept) which describes the considered concept the best due to a language bias, background knowledge and dataset.*

'The best' in above definition means that the hypothesis covers the dataset, and is not in the contradiction with the background knowledge and holds true with the language bias.

Methods presented in this work, as well as most of methods of KD and DM, are inductive methods. It means that knowledge is formed as a generalization of observations.

### 1.3.4 Identification

In the topic of this work appeared a term *identification*. Because in the spoken language 'identification' has the same meaning as 'recognition' which is not true in computer science, so to avoid any misunderstandings a good definition is needed. *The system identification* [3, 33, 55] is a field of science which includes methods for modelling systems or objects. These methods are based on statistics, mathematical approximation and numerical methods. There are two main problems in system identification:

- When a considered system is described by a model known in accuracy to pa-
rameters, then it is a problem of parameter identification.

- When there is a known class of models and the best model has to be chosen, then it is a problem of optimal model selection.

In this work we assume that there is a model of a relation describing a numeric attributes of the system (given by an expert) known in accuracy to parameters. Depend on background knowledge and dataset we can use parameter estimation or optimal model selection algorithms.

1.3.5 Main goal of the work

Very useful and convenient way of representing knowledge is a mathematical modelling using algebraic formulas \([3, 33]\) or logic formulas \([5, 7, 8, 13, 14, 20, 25, 26, 29, 30, 34, 39, 48, 65]\). However, mainly the knowledge about an object is described by a functional representation which makes it easy to process, aggregate or analyse. Nevertheless, very often the knowledge about the object cannot be represented by classical functional descriptions, because of its properties or its uncertainty (non-deterministic systems). Experts or systems designers cannot exactly describe considering system, because e.g. they are not able to say anything about noise signals or external disturbances acting on the system. Therefore there is a need to develop new methods for knowledge representation. One of this approach is a relational knowledge description \([4] - [13], [26, 42, 50, 51, 52, 53, 54, 56, 57, 58, 60, 61]\).

Generally could be said that there exists a relation describing knowledge object. This relation could be in such forms: an equation of a circle or set of inequalities, etc. However in many cases expert cannot give exact relational knowledge representation. He or she could formulate problem in accuracy to unknown parameters. Then, having data (observations), could be given some algorithm which identifies those unknown parameters. Problem of a parameter identification in a relational system (so called relation identification) is a generalized problem of an identification for functional knowledge representation.

In this work methods of the relation identification are presented. This methods are formulated to use them in the knowledge extraction from dataset. As an output of the identification method and a general knowledge discovery method there is a property of the system (relation) described by rule (set of rules) written in first-order logic formulas.
Such extracted knowledge from dataset, written in logical formulas, can be later used in decision making problem or analysis problem [4], [6] - [12], [59] by applying logic-algebraic method (LAM) [4] - [12], [59] which is a generalization of the reasoning method. Until now no relation identification methods for discovering knowledge for real expert systems were proposed. Therefore methods presented in this work combining with logic-algebraic method can be used for designing complex expert systems.

1.4 Research questions and outcomes

1.4.1 Research questions

In this work such research questions are formulated:

Q1: How could the knowledge object be treated as the relational system?
Q2: What are knowledge discovery problems for relational systems (relational knowledge objects)?

1. How is the relational system defined?
2. What is the general problem statement for knowledge discovery in the relational system?
3. What is the general procedure of knowledge discovery in the relational system?
4. What kinds of relational models could be used?
5. What kind of knowledge discovery problems could be stated?

1.4.2 Outcomes

After answering the research questions following outcomes are expected:

- description of an idea of knowledge discovery using relational identification methods for different kinds of models and problem statements;
- first coherent work describing idea of knowledge discovery using relational identification methods;
• formulation of identification methods for different relational models which could be used in expert systems;

• comparison of presented approach with well known data mining methods on the example of well-known dataset;

• application of presented approach on the real dataset.

1.5 Work contents

The work is organized as follows:

Chapter 1 provides an introduction to the topic of the thesis describing the problem background, basic definitions, research questions and outcomes.

Chapter 2 gives crucial definitions and the general problem statement and the idea of knowledge discovery method.

In the Chapter 3 two methods for relation identification are presented and modified coverage knowledge discovery method is shown.

In Chapter 4 and Chapter 5 relation identification problems for parameter estimation and optimal model selection, respectively, are formulated.

Chapter 6 provides general approach to relation identification with additional uncertain description.

In the Chapter 7 three applications of relational knowledge representation are presented: economic case, biomedical case and benchmark dataset.

Finally, in Chapter 8 overall conclusions are conveyed along with perspectives of future work.
Chapter 2

Basic definitions and problem statement

In this chapter a new approach to the knowledge extraction is presented which is associated with a relational description of a knowledge object/system (a relational object/system). In this chapter a relational object is described and other useful definitions are formulated, e.g. knowledge representation, uncertainty. All of ideas are showed in connection with knowledge discovery. Furthermore, a general idea of an algorithm for knowledge discovery is proposed and a general problem of relation identification is stated.

2.1 Relational object

Let us consider a knowledge object (system) in which an input, \( x \in X \subseteq \mathbb{R}^S \), which is a vector \( x = (x^{(1)}, x^{(2)}, \ldots, x^{(S)}) \), and an output, \( y \in Y \subseteq \mathbb{R}^L \), which is a vector \( y = (y^{(1)}, y^{(2)}, \ldots, y^{(L)}) \), can be distinguished. Moreover, a physical quantity about this object can be described by a relation [9] - [12], [44]:

\[
x \rho \ y \triangleq R(x, y) \subseteq X \times Y
\]

which is a set of all possible pairs \((x, y)\) in the Cartesian product \(X \times Y\) which may appear in the object.

**Definition 2.1.** (Relational object/system) The relational object/system is described by set of properties (predicates) concerning \((x, y)\); \(R(x, y)\) denotes the set of all pairs \((x, y)\) for which properties are satisfied,

\[
R(x, y) = \{(x, y) \in X \times Y : w[\Phi(x, y)] = 1\} \tag{2.1}
\]
where:
\( w[\cdot] \) – a logic value, \( w[\cdot] \in \{0, 1\} \);
\( \Phi(x, y) \) – a complex predicate describing the object that is compound of \( M \) single predicates, \( \{ \phi_1(x, y), \phi_2(x, y), \ldots, \phi_M(x, y) \} \), for which following expression holds true:
\[
    w[\Phi(x, y)] = \bigwedge_{m=1}^{M} w[\phi_m(x, y)] = 1.
\]

**Remark 2.1.** The single predicate, \( \phi_m \), is built of logic operations \( \land \) – and, \( \lor \) – or, \( \rightarrow \) – 'if... then...', \( \neg \) – negation, and can concern the input property (e.g. \( \phi_x = "x > 20" \)), the output property (e.g. \( \phi_y = "y = 1" \)), or input-output property (e.g. \( \phi = "y > 2x + 1" \)).

**Remark 2.2.** To simplify further notation following expression is used:
\[
    \lambda_i \triangleq w[\phi_i(x, y)] = \begin{cases} 
        1, & \text{if } i^{th} \text{ property holds true in logical sense} \\
        0, & \text{otherwise}
    \end{cases}
\]

Now it is possible to define a relational knowledge representation about the object/system.

**Definition 2.2.** *(Relational Knowledge Representation)* The relational knowledge representation (RKR) is following tuple

\[
    RKR = \langle B, \mathcal{L}, X, Y, \Phi \rangle \tag{2.2}
\]

where:
\( B \) – a background knowledge;
\( \mathcal{L} \) – a language bias;
\( X \) – an input space;
\( Y \) – an output space;
\( \Phi \) – a complex predicate about the object.
Very important is to point out what can be included in the background knowledge:

- knowledge about attributes like: names (labels), information if attribute is an input, output or input/output (attribute can be treated as input or output as well);
- additional predicates (e.g. describing the property about the object);
- statements about the object (e.g. describing the physical quantity);
- other (e.g. description of the system structure, description of the connection of the considered object with other system's components, additional uncertainty descriptions).

Without the background knowledge a relation cannot be treated as a knowledge in this sense which was given in the chapter 1. Because for example following relation, \( x \in \mathbb{R}^1, y \in \mathbb{R}^1 \):

\[
R(x, y) = \{(x, y) \in \mathbb{R}^1 \times \mathbb{R}^1 : w\left[ (x > 0) \land (y > 0) \land (y = 2 \cdot x) \right] = 1 \},
\]

without any context (no knowledge about labels and statements about the object), is only a mathematical expression and is an information (ordered data). And can be treated as a set of all even numbers rather than, for example, a concept about working of an engine.

We assume that the background knowledge is formulated before the knowledge discovery process is applied and can be given by an expert or some other credible source.

### 2.2 Partial relation

It was said that relational system has a complex predicate which contains of \( M \) single predicates. Thus we can distinguish a part of the relation which is described by a single predicate, \( \phi_m(x, y) \). By partial relation we understand following expression:

\[
\tilde{R}_m(x, y) = \{(x, y) \in X \times Y : w[\phi_m(x, y)] = 1 \}.
\]

where:
\( \phi_m(x, y) \) is one of the simple predicate which is a partial description of the relational system.
Partial relation can be called also an association among attributes.

In inductive logic programming or relational data-mining (data-mining in relational databases) [18, 30] it is assumed that most of the partial relations are known (each association is represented by a table). Then more complicated predicates are formulated using partial relations. In the approach presented in this paper this partial relations are unknown.

2.3 Relational model

In real-life situations it is often impossible to get a full or even partial description of an object (a system). Therefore some approach to receive this knowledge should be proposed. To achieve that goal we assume that there is a model concerning the system or we propose some model of the system.

**Definition 2.3. (Model of the relational object / relational model)** The relational model is a mathematical description known in accuracy to unknown parameters $a \in A$,

$$R_a(x, y) \triangleq R(x, y; a) = \{(x, y) \in X \times Y : w[F(x, y; a)] = 1\} \quad (2.4)$$

where:
- $a$ – unknown parameters, $a \in A$;
- $x$ – input;
- $y$ – output;
- $F(x, y; a)$ – proposed description of system’s property.

**Remark 2.3.** Proposed description of partial relation is denoted as $\tilde{F}_m(x, y; a)$.

If it is possible to give any structure of a model, then the proposed description $F(x, y; a)$ is a part of the background knowledge $B$, as well as the set of unknown parameters $A \subseteq \mathbb{R}^k$.

However, proposing the complex property with all details is rather a difficult task or even impossible, so that a special methods have to be proposed. In this work we say that only a model about the numeric attributes can be proposed and at least one of these attributes is input- or input/output-like, and at least one is output- or input/output-like.
Then, if above conditions hold true, we say that proposed model is a partial description of considered relation. And in compound with other procedure can produce more detailed description.

Without knowing the exact form of relation property $\Phi$, we can formulate new relational knowledge representation.

**Definition 2.4. (Relational Knowledge Representation with model)** The Relational Knowledge Representation with a model is as follows:

$$RKR = \langle B, \mathcal{L}, X, Y, A, R_a \rangle$$

where:
- $B$ – a background knowledge;
- $\mathcal{L}$ – a language bias;
- $X$ – an input space;
- $Y$ – an output space;
- $A$ – a space of unknown parameters;
- $R_a$ – a model of relation (part of relation).

### 2.4 Dataset description

Very important issue is to describe dataset. Dataset or a collection of examples about the concept (object) is viewed as a table with attributes’ values in columns, and with instances in rows. An example dataset is shown in the table 2.1 where $x^{(s)}$ is $s^{th}$ input, $s = 1, 2, \ldots, S$, $y^{(l)}$ is $l^{th}$ output, $l = 1, 2, \ldots, L$, and $\varphi_x(n, s)$ is an information function which for $n^{th}$ instance gives a value of $s^{th}$ input, and $\varphi_y(n, l)$ is an information function for output. Then the dataset could be called information system [40].
No.  | $x^{(1)}$  | $x^{(2)}$  | ... | $x^{(S)}$ | $y^{(1)}$ | $y^{(2)}$  | ... | $y^{(L)}$
---|---|---|---|---|---|---|---|---
1   | $\varrho_x(1,1)$ | $\varrho_x(1,2)$ | ... | $\varrho_x(1,S)$ | $\varrho_y(1,1)$ | $\varrho_y(1,2)$ | ... | $\varrho_y(1,L)$
2   | $\varrho_x(2,1)$ | $\varrho_x(2,2)$ | ... | $\varrho_x(2,S)$ | $\varrho_y(2,1)$ | $\varrho_y(2,2)$ | ... | $\varrho_y(2,L)$
... | ... | ... | ... | ... | ... | ... | ... | ... | ...
N   | $\varrho_x(N,1)$ | $\varrho_x(N,2)$ | ... | $\varrho_x(N,S)$ | $\varrho_y(N,1)$ | $\varrho_y(N,2)$ | ... | $\varrho_y(N,L)$

Table 2.1: Example of a dataset.

Furthermore, we assume that there is no \textit{a priori} information about the system and connections between input and output. If there is possibility to distinguish connections between input and output (special knowledge in $\mathcal{B}$), then the system can be decomposed to a simpler (or simple – consisting only of one predicate) objects and each object can be determined separately (like in multi-relational data mining \cite{18}). Otherwise the system has to be considered as a whole.

Remark 2.4. Because in classification theory \cite{28} distinguished output attributes are called \textit{classes}, therefore it will be made so in this work if it makes theory more clear to explain.

Moreover, we assume that there are no missing values or semantically the same values (e.g. "Pepsi" and "Pepsi Cola"). In general, if it is not said otherwise, we assume that dataset is pre-processed. Nevertheless, a noise is possible in data.

Referencing to dataset in form of a table is very unpractical and difficult, therefore we say that there is a sequence of observations (examples, instances), $(x_n, y_n, \lambda_n)$, $n = 1, 2, \ldots, N$, that we write as follows:

$$R_N = \{(x_n, y_n, \lambda_n) \in X \times Y \times \Lambda : n = 1, 2, \ldots, N\}$$

(2.6)

where:

$x_n$ – input values of $n^{th}$ instance;

$y_n$ – output values of $n^{th}$ instance;

$\lambda_n$ – belonging of $n^{th}$ instance to the relation, $\lambda_n \in \Lambda$;

In general $\Lambda$ represents a level of belonging of $n^{th}$ instance to the relation, $\Lambda = [0, 1]$, or in particular – $\Lambda = \{0, 1\}$ which has more intuitive meaning. If an instance belongs to the relation, then $\lambda_n = 1$, or if does not, then $\lambda_n = 0$. In our further considerations we investigate only the second case. Thus we can divide $R_N$ into
two subsets:

\[ R_N = R_{N,0} \cup R_{N,1} \]

where:

\[ R_{N,0} = \{(x_n, y_n, \lambda_n) \in R_N : \lambda_n = 0\} \] – so called negative examples;
\[ R_{N,1} = \{(x_n, y_n, \lambda_n) \in R_N : \lambda_n = 1\} \] – so called positive examples.

Obviously for those sets \( R_{N,0} \cap R_{N,1} = \emptyset \) holds true.

For further simplicity let us introduce following sets of indexes:

- \( E_N = \{n = 1, 2, \ldots, N\} \) – set of indexes of all examples;
- \( E_{N,0} = \{n : \lambda_n = 0\} \) – set of indexes of negative examples;
- \( E_{N,1} = \{n : \lambda_n = 1\} \) – set of indexes of positive examples;
- \( E_B = \{n : B\} \) – set of indexes of examples which are not in contradiction with background knowledge;
- \( E_{N,B} = \{n : x = b\} \), set of indexes of examples where \( b \) is a set of chosen values of input used in proposed description body.

In above set \( E_{N,B} \) a new term occurred, the proposed description body. We will look for a description of a relation in a form of implication (or conjunction of implications),

\[ \phi : B \Rightarrow C \]

where:

- \( B \) – conditions (body);
- \( C \) – conclusions.

Therefore \( E_{N,B} \) is a set of indexes of instances for which \( B : "x = b" \).

### 2.5 Uncertainty

The uncertainty plays a main role in the information science and information engineering [6, 9, 10, 40]. By the uncertainty we understand a situation when about some object (e.g. a quantity, a decision) there is no complete information available. In other words – there is no certainty, confidence about a behaviour of
the object. However, there is some pattern in the behaviour which can be formalised by a mathematical formulas. There are several uncertainty descriptions, but the most common are following:

- random variables \([9, 10, 23]\);
- fuzzy sets, fuzzy numbers, fuzzy relations \([9, 10, 16]\);
- uncertain variables \([9, 10, 59]\);
- rough sets \([40, 41]\);
- others (e.g. soft variables \([10]\), shadowed sets \([14]\)).

In this work some methods for knowledge discovery with additional knowledge about uncertainty is presented.

### 2.6 Dissimilarity measure

In our considerations we look also for fitting our data to the model. Thus, in other words, we are interested in counting how much our relation represented by dataset \(R_N\) is not similar to the model \(R_a\). Therefore for given \((x_n, y_n) \in X \times Y\) and \(a \in A\) we can count dissimilarity between \(R_N\) and \(R_a\) \([50]\):

\[
\tilde{q}[R_N, R_a, (x_n, y_n)] \triangleq q[\lambda_n, \bar{\lambda}_n]
\]  

(2.7)

where:
- \(\tilde{q}\) – dissimilarity measure;
- \(\lambda_n\) – logic value of \(F(x_n, y_n; a)\);
- \(q\) – is a function \(q: \{0, 1\} \times \{0, 1\} \rightarrow \mathbb{R}^+\) for which \(q[\lambda, \bar{\lambda}] = 0 \iff \lambda = \bar{\lambda}\) holds true.

For example, for given \(a \in A\), dissimilarity measure can be as follows:

\[
\tilde{q}[R_N, R_a, (x_n, y_n)] = \begin{cases} 
0 & \text{if } (x_n, y_n) \in R_N \land (x_n, y_n) \in R_a \\
q_1[(x_n, y_n), R_N] & \text{if } (x_n, y_n) \notin R_N \land (x_n, y_n) \in R_a \\
q_2[(x_n, y_n), R_a] & \text{if } (x_n, y_n) \in R_N \land (x_n, y_n) \notin R_a \\
0 & \text{if } (x_n, y_n) \notin R_N \land (x_n, y_n) \notin R_a 
\end{cases}
\]  

(2.8)

or

\[
q[\lambda_n, \bar{\lambda}_n] = \begin{cases} 
0 & \text{if } \lambda_n = 1 \land \bar{\lambda}_n = 1 \\
q_1[(x_n, y_n), R_N] & \text{if } \lambda_n = 0 \land \bar{\lambda}_n = 1 \\
q_2[(x_n, y_n), R_a] & \text{if } \lambda_n = 1 \land \bar{\lambda}_n = 0 \\
0 & \text{if } \lambda_n = 0 \land \bar{\lambda}_n = 0 
\end{cases}
\]  

(2.9)
where:
\( \varrho_1, \varrho_2 \) – measures of distance (metrics);

Metrics \( \varrho_1, \varrho_2 \) can be following:

\[
\varrho_1[(x, y), R_N] = \inf_{(u, v) \in R_N} \rho[(x, y), (u, v)],
\]

\[
\varrho_2[(x, y), R_a] = \inf_{(u, v) \in R_a} \rho[(x, y), (u, v)].
\]

Or, in particular, a zero-one metric,

\[
\varrho[(x, y), (u, v)] = \begin{cases} 
0, & \text{if } (x, y) = (u, v) \\
1, & \text{otherwise}
\end{cases}
\] (2.10)

### 2.7 Quality criteria

In analogy to classical identification methods, it can be formulated a quality criterion, \( Q(\cdot) \), for assessing parameter's values.

\[
Q(a) = \varrho_N \{ q[\lambda_n, \bar{\lambda}_n] \}
\] (2.11)

where:

\( \varrho_N \) is a dissimilarity measure between a relational model, \( R_a \), and dataset, \( R_N \) (e.g. \( \varrho_N = \min \) or \( \varrho_N = \sum \)).

Let us formulate a definition of optimal model.

**Definition 2.5.** (Optimal model) We say that relation \( R_a \) is an optimal model for unknown relation \( R \) having data \( (x_n, y_n), \lambda_n, n = 1, 2, \ldots, N \), for which value \( a = a_N \) minimizes a quality criterion,

\[
Q(a_N) = \min_{a \in A} Q(a).
\]

**Remark 2.5.** Very useful from the practical point of view and very intuitive is following criterion:

\[
Q(a) = \sum_{n \in \mathcal{N}_1(a)} 1 + \sum_{n \in \mathcal{N}_2(a)} 1
\] (2.12)

where:

\( \mathcal{N}_1(a) = \{ n \in [1, N] : \lambda_n = 1 \wedge \bar{\lambda}_n = 0 \} \),
\[ N_2(a) = \{ n \in \overline{1,N} : \lambda_n = 0 \land \bar{\lambda}_n = 1 \}, \]
or
\[ Q(a) = card\{N_1(a)\} + card\{N_2(a)\}. \quad (2.13) \]

where:

\( card\{\cdot\} \) – cardinality of a set.

This criterion can be called a measure of misclassification.

It can be obtained from putting zero-one metric (2.10) into (2.8) [50]. And it can be explained as a number of negative examples included by relation model plus number of positive examples outside the model.

### 2.8 Idea of knowledge discovery method

Knowledge discovery process (KDP) is a nontrivial process of processing and validating data from which information are created; conclusions driven from information formulate knowledge [19, 20]. Information and knowledge are written in some knowledge representations, not necessarily the same.

![Knowledge discovery process](source: [19]).

KDP process (figure 2.1) consists of following steps:

1. Data selection.
2. Data preprocessing (e.g. data cleaning, data formats standardization, data integration).
3. Data transformation (e.g. preparing data for data mining methods).

4. Data mining (discovering associations, patterns, information in data).

5. Drawing conclusions from information (e.g. DM results interpretation or evaluation).

More detailed descriptions of KDP could be found in the literature (e.g. [14, 19]).

In this work it is assumed that dataset is selected, preprocessed and transformed. So it can be said that in our considerations the KDP starts from the level of transformed data.

2.9 General problem statement

As it was mentioned in chapter 1, we use induction as a reasoning method. Therefore, general problem statement is as follows:

Problem Statement 2.1. General problem statement

GIVEN:

- Selected and pre-processed dataset $R_N$.
- Background knowledge $B$.
- Language bias $L$.
- Relational knowledge representation.

FIND:

- Description $F$.

To solve above problem we use decomposition method of dataset into subsets of data (partial relations). Final relation is a merged description of associations. Idea of knowledge discovery method is presented in the figure 2.2.
Figure 2.2: Idea of knowledge discovery method.
Chapter 3

Relation identification methods as knowledge discovery methods

In this chapter different problem statements and methods for knowledge discovery are presented. First two methods for relation identification are formulated and then general procedure for knowledge discovery.

3.1 Preliminaries

Before formulating knowledge discovery methods some ideas have to be introduced.

Generally in inductive way of reasoning it is important to find such knowledge description which covers all positive examples and exclude all negative examples. In analogy to pattern recognition, it is an equivalent of classification accuracy at the level of 100%. There are two concepts:

- **Completeness** – how many positive examples are included by the knowledge description.
- **Consistency** – how many negative examples are included by the knowledge description.

Moreover, we assume that data are credible, which means that there is no noise or missing values. Therefore we are interested in finding such description which is fully complete and consistence.
To express completeness and consistency in formal way, for partial relations in form of implications, two measures can be proposed:

- **Coverage**:

  \[ coverage(F \cup B, E_N) = \frac{\text{card}\{ E_{N,B} \cap E_{N,1} \cap E_B \}}{\text{card}\{ E_{N,1} \}} \]  
  \[ (3.1) \]

- **Accuracy**:

  \[ accuracy(F \cup B, E_N) = \frac{\text{card}\{ E_{N,B} \cap E_{N,1} \cap E_B \}}{\text{card}\{ E_{N,B} \}} \]  
  \[ (3.2) \]

where \( \text{card}\{ \cdot \} \) describes cardinality of a set.

Obviously coverage describes completeness, and accuracy – consistency. An ideal situation is when \( coverage = 1 \), and \( accuracy = 1 \).

**Remark 3.1.** If we take a performance index as:

\[ Q(a) = \sum_{n \in \mathcal{N}_1(a)} 1 + \sum_{n \in \mathcal{N}_2(a)} 1 \]

where:

\[ \mathcal{N}_1(a) = \{ n \in \overline{1,N} : \lambda_n = 1 \land \bar{\lambda}_n = 0 \}, \]

\[ \mathcal{N}_2(a) = \{ n \in \overline{1,N} : \lambda_n = 0 \land \bar{\lambda}_n = 1 \}. \]

and normalizing both sums (\(| \cdot |\) is cardinality of a set),

\[ Q(a) = \frac{\frac{|\mathcal{N}_1(a)|}{|\mathcal{N}_1(a) \cup \mathcal{N}_2(a)|} + \frac{|\mathcal{N}_2(a)|}{|\mathcal{N}_1(a) \cup \mathcal{N}_2(a)|}}{2}, \]

then we can express this criterion in terms of \( coverage \) and \( accuracy \):

\[ Q(a) = \frac{(1 - coverage) + (1 - accuracy)}{2} = 1 - \frac{1}{2}(coverage + accuracy). \]

For further simplicity of notation let us introduce definition of **simple formula**.

**Definition 3.1.** *(Simple formula)* By simple formula we understand an attribute which equals to some value. Simple formula is denoted by \( \alpha \).
Remark 3.2. By $\alpha_x$ we denote input, $\alpha_y$ - output, $\alpha_w$ – additional features.

For example: $\alpha_x = "\text{temperature} > 37\text{"}$, $\alpha_w = "\text{ID} = 12345\text{"}$, $\alpha_y = "\text{illness} = \text{flu}\text{"}$.

3.2 Knowledge discovery methods

Now we can formulate general methods for relation identification. We can distinguish three situations:

1. There is numeric dependency between input and output and only positive examples are known.
2. There is numeric dependency between input and output and positive and negative examples are known.
3. Output is discrete and its values belong to a finite set.

In the first and second case methods are very similar but in the second one some criterion has to be formulated. Moreover, we are interested in finding the most exact descriptions for which coverage and accuracy are highest so we look also for additional formulas, $\alpha_w$. Both methods are called identification methods because we propose some model of the object.

In the third case we discover knowledge in form of implications (clauses) using modified covering strategy [34, 65]. This method includes relation identification method if there are some numeric attributes.

In each case we want to get highest coverage with highest accuracy. Therefore any criterion has to fulfill this condition.

3.2.1 Relation identification method for positive examples

Let us introduce problem for positive examples only.

Problem Statement 3.1. Relation identification problem for positive examples

GIVEN:

- Only positive instances are available, $R_N \equiv R_{N,1}$. 
• Output is numeric.

• Input is numeric, \(x_{\text{num}} \in X_{\text{num}}\), and nominal, \(x_{\text{nom}} \in X_{\text{nom}}\); \(X = X_{\text{num}} \times X_{\text{nom}}\).

• Partial relational models, \(\bar{R}_a(x_{\text{num}}, y)\).

**FIND:**

• Find a relation description \(F(x, y; a)\) which covers \(R_N\) completely, in other words – a set \(D_a(N) \subseteq A\) such that

\[
a \in D_a(N) \Rightarrow \forall_{n=1,\ldots,N} (x_n, y_n) \in R_a(x, y)
\]

However, in many practical situations we are interested in achieving only one vector of parameters (only one relation), therefore we look for a boundary of \(D_a(N)\), \(\Delta_a(N)\), which is a singleton, \(\Delta_a(N) = \{a_N\}\).

Moreover, it is not allowed to create implications or logic expressions in which one nominal attribute can equal to several values. It could be easily illustrated on a simple example: "illness = YES" \(\land\) "illness = NO" \(\Rightarrow\) "treatment = YES". Both values occur in the same moment that makes no sense at all. Therefore in our methods if one nominal attribute is used in the logic expression, then other values of this attribute can be added only as a disjunction.

**Remark 3.3.** Set of all possible values of nominal \(i^{th}\) attribute, \(x_i\), is denoted by

\[
X_i = \{j_i : x_i = j_i\}.
\]

**Procedure 3.1. Relation identification method for positive examples**

**INPUT:** \(R_{N,1}, X_{\text{num}}, X_{\text{nom}},\) output is numeric, \(\bar{R}_a(x_{\text{num}}, y)\).

**START:** \(F(x, y; a) = \emptyset\).

**STEP 1**
Take the partial relational model \(\bar{R}_a(x_{\text{num}}, y)\).
CHAPTER 3. METHODS

27

STEP 2
Calculate a set \( D_a(N) \subseteq A \) such that
\[
a \in D_a(N) \Rightarrow \forall_{n=1,...,N}(x_n, y_n) \in \tilde{R}_a(x_{nom}, y)
\]
(In other words: for given \( a \in D_a(N) \) \( \text{coverage}(\tilde{F}, E_N) = 1 \)).

STEP 3
Calculate a boundary \( \Delta_a(N) \) of \( D_a(N) \).

STEP 4
Create a model, \( \tilde{R}_{a_N} \), it means partial description \( \tilde{F}(x, y; a_N) \).
Set \( F(x, y; a_N) = \tilde{F}(x, y; a_N) \).
If \( X_{nom} \equiv \emptyset \), then print out \( F(x, y; a_N) \) and STOP.
Else \( i = 1 \).

STEP 5
If \( i > |X_{nom}| \), then print out \( F(x, y; a_N) = F(x, y; a_N) \land B \) and STOP.
Else take \( x_{nom,i} \) and calculate coverage for all possible values of attribute \( x_{nom,i} \), \( j_i \in X_{nom,i} \).
\[
\forall_{j_i \in X_{nom,i}} \text{coverage} ("x_{nom,i} = j_i", E_N)
\]
Add simple formula "\( x_{nom,i} = j_i^n \)" for which \( \text{coverage} = 1 \), as a conjunction with \( F(x, y; a_N) \),
\[
F(x, y; a_N) = F(x, y; a_N) \land "x_{nom,i} = j_i^n".
\]
If there are more simple formula which has \( \text{coverage} = 1 \), then add them as a conjunction of disjunctions
\[
F(x, y; a_N) = F(x, y; a_N) \land ("x_{nom,i} = j_i^1 \lor x_{nom,i} = j_i^2 \lor \cdots \lor x_{nom,i} = j_i^k").
\]
Set \( i = i + 1 \) and GOTO STEP 5.

OUTPUT: \( F(x, y; a_N) \).
3.2.2 Relation identification method for positive and negative examples

Let us introduce problem for positive and negative examples.

*Problem Statement 3.2. Relation identification problem for positive and negative examples*

**GIVEN:**
- Both positive and negative instances are available, $R_N$.
- Output is numeric.
- Input is numeric, $x_{num} \in X_{num}$, and nominal, $x_{nom} \in X_{nom}$, $X = X_{num} \times X_{nom}$.
- Partial relational models, $\tilde{R}_a(x_{num}, y)$.
- A performance criterion, $Q(\cdot)$.

**FIND:**
- Find a relation description $F(x, y; a)$ which covers as much as possible of $R_N$ with highest accuracy, in other words – calculate a set $D_a(N) \subseteq A$ so that

$$a^* \in D_a(N) \Rightarrow Q(a^*) = \min_{a \in A} Q(a).$$

**Procedure 3.2. Relation identification method for positive and negative examples**

**INPUT:** $R_N$, $X_{num}$, $X_{nom}$, $y$ is numeric, $\tilde{R}_a(x_{num}, y)$, $Q(\cdot)$.

**START:** $F(x, y; a) = \emptyset$.

**STEP 1**
Take the partial relational model $\tilde{R}_a(x_{num}, y)$.

**STEP 2**
Calculate a set $D_a(N) \subseteq A$ using criterion $Q$.
CHAPTER 3. METHODS

STEP 3
Calculate a boundary $\Delta_a(N)$ of $\mathcal{D}_a(N)$.

STEP 4
Create a model, $\tilde{R}_{a_N}$, it means partial description $\tilde{F}(x, y; a_N)$.
Set $F(x, y; a_N) = \tilde{F}(x, y; a_N)$ and calculate coverage and accuracy.
If $X_{\text{nom}} = \emptyset$, then print out $F(x, y; a_N)$ and STOP.
Else $i = 1$ and GOTO STEP 5.

STEP 5
If $i > |X_{\text{nom}}|$, then print out $F(x, y; a_N) = F(x, y; a_N) \land B$ and STOP.
Else take $x_{\text{nom},i}$ and set $k = 1$.

STEP 6
If $k > |X_{\text{nom},i}|$, then $i = i + 1$ and GOTO STEP 5.
Add simple formula $\alpha_{x_{\text{nom},i}} = "x_{\text{nom},i} = j^k_i"$ to the description of relation as a conjunction,
\[ F(x, y; a_N) = F(x, y; a_N) \land "x_{\text{nom},i} = j^k_i". \]
Calculate coverage and accuracy for the new description.
If coverage and accuracy have risen or are the same as previous, then remove simple formula, $q = 1$ and GOTO STEP 7.
Else remove formula "$x_{\text{nom},i} = j^k_i"$ from description, $k = k + 1$ and GOTO STEP 6.

STEP 7
If $k + q > |X_{\text{nom},i}|$, then $F(x, y; a_N) \land \alpha_{x_{\text{nom},i}}$, $i = i + 1$ and GOTO STEP 5.
Else broaden simple formula $\alpha_{x_{\text{nom},i}} = "x_{\text{nom},i} = j^k_i"$ by adding one of the rest possible attribute values of $i^{th}$ attribute, "$\alpha_{x_{\text{nom},i}} = \alpha_{x_{\text{nom},i}} \lor "x_{\text{nom},i} = j^{k+q}_i"$. Calculate coverage and accuracy and if they have risen or are the same as previous, $q = q + 1$ and GOTO STEP 7.
Otherwise remove last added alternative from simple formula, $q = q + 1$ and GOTO STEP 7.

OUTPUT: $F(x, y; a_N)$. 
3.2.3 Modified coverage knowledge discovery method

In a case when output is nominal and/or there is a finite set of possible output values then relation identification methods are not valid. Thus more general procedure has to be proposed.

In this point a modified coverage strategy is presented. The main difference between standard coverage strategy and proposed approach is that partial relation identification method is included. Besides, both measures, coverage and accuracy, play crucial role in finding relational description in form of logical expressions.

This approach, called modified coverage method, is based on finding rules (clauses) which describe each output. It means that we have no clue how complex predicate should look like and we "guess" it by formulating logical expressions for outputs in form of implication (clause).

First we decompose dataset due to output values (decomposition into partial relations). Second we create partial descriptions. And finally we merge all descriptions as a conjunction. This method is the same as it was showed in the figure 2.2.

Problem Statement 3.3. Relation discovery problem

GIVEN:

- Dataset, $R_N$.
- Output is nominal and a possible set of classes is finite, $y \in \{y_1, y_2, \ldots, y_L\}$.
- Input is numeric, $x_{num} \in X_{num}$, and nominal, $x_{nom} \in X_{nom}$; $X = X_{num} \times X_{nom}$.
- Partial relational model, $\tilde{R}_a(x_{num}, y)$.
- A performance criterion, $Q(\cdot)$.

FIND:

- Find a relation description $F(x, y; a)$ which covers as much as possible of $R_N$ with highest accuracy.
Procedure 3.3. Relation discovery method

**INPUT:** $R_N$, $X_{num}$, $X_{nom}$, output is nominal and $y \in \{y_1, y_2, \ldots, y_L\}$, $\tilde{R}_a(x_{num}, y)$, $Q(\cdot)$.

**START:** $F(x, y; a) = \emptyset$.

**STEP 1**
If $X_{num} \neq \emptyset$, then take the numeric part of input. Set part of them as ”input”, $x_{num,in} \equiv \bar{x}$, and rest as ”output”, $x_{num,out} \equiv \bar{y}$.
Else $k = 1$ and GOTO STEP 3.

**STEP 2**
If $K = 1$, then apply procedure 3.1, $\Psi_1$, with $\bar{X}_{nom} = \emptyset$. Result is $\Psi_1(R_N, A) = a_N$ and partial relation $\tilde{R}_{a_N}(\bar{x}, \bar{y})$.
Else:
For each class $y_l$, $l = 1, 2, \ldots, L$, apply procedure 3.2, $\Psi_2$, with $\bar{X}_{nom} \equiv \emptyset$ and $\lambda_n = 1$ if class of $n^{th}$ instance $y_n = y_l$ and $\lambda_n = 0$ - otherwise. Result is $\Psi_2(R^{(l)}_N, A) = a_N$ and partial relation $\tilde{R}_{a_N}^{(l)}(\bar{x}, \bar{y})$.
Set $k = 1$.

**STEP 3**
If $l > L$, then GOTO STEP 6.
Create a partial relation description $\tilde{F}_l(x, y; a_N)$ in form:

$$(B_l \equiv \emptyset) \Rightarrow y_l.$$ 

**STEP 4**
For class $y_l$ calculate coverage and accuracy for all nominal attributes with different values and for distinct partial relation,

$$\tilde{R}_{a_N}^{(l)} = \tilde{R}_{a_N}^{(l)} \setminus \bigcup_{p=1, \ldots, L, p \neq l} \tilde{R}_{a_N}^{(p)}.$$ 

Add to the body this attribute value or partial relation which has the highest coverage. To the chosen attribute value or partial relation add as a disjunction all simple formula for which accuracy = 1.
Calculate coverage and accuracy of $\tilde{F}_l$. 

**CHAPTER 3. METHODS**
**CHAPTER 3. METHODS**

**STEP 5**
If there is no simple formula left, then $\tilde{F}_l(x, y; a_N) = "B_l \Rightarrow y_l", \ l = l + 1$ and GOTO STEP 3.

Else take not yet added simple formula, $\alpha$, with highest coverage and add it to the body of description as a conjunction, $B_l \land \alpha$. Calculate coverage and accuracy.

If coverage and accuracy have risen or are the same as previous, then GOTO STEP 5.

Else remove formula $\alpha$ and GOTO STEP 5.

**STEP 6**
Bound all implications as a conjunction, $l = 1, 2, \ldots, L$, and create final description:

$$F(x, y; a_N) = \tilde{F}_1 \land \tilde{F}_2 \land \cdots \land \tilde{F}_L \land \mathcal{B}.$$

**OUTPUT:** $F(x, y; a_N)$.

**Remark 3.4.** For $L = 2$ it is enough to determine a partial relation for positive examples $R_{N,1}, \bar{R}_{a_N}$. Then for negative examples $R_{N,0} \neg \bar{R}_{a_N}$ can be taken as a partial relation.
Chapter 4

Relation identification problem for parameter estimation

In this chapter several algorithms for parameter estimation problem are presented. Moreover, only positive examples with only numeric input and numeric output are considered.

4.1 Problem outline

In Chapter 3 a general method for relation identification for positive examples was presented. However, for different relation models different algorithms have to be proposed [5, 50]. Moreover, in many cases special numerical methods have to be formulated because analytical approach is too complicated or even impossible [5, 50, 55].

4.2 Problem formulation

Problem formulation is the same as it was presented in Problem statement 3.1. Let us briefly remind this problem.

There are given only positive examples, $R_{N1}$, input and output are numeric and a relational model is known with accuracy to unknown parameters $a \in A$. We are interested in finding such parameters values, $D_a(N)$, for which following implication is fulfilled:

$$a \in D_a(N) \Rightarrow \forall_{n=1,...,N} (x_n, y_n) \in R_a(x, y).$$
In other words we want to find such relation that covers all examples.

Besides, from practical point of view we look only for a boundary of \( D_a(N) \), \( \Delta_a(N) \), which is a singleton, \( \Delta_a(N) = \{ a_N \} \). Mainly we choose only this parameters for which volume of the relation (in the sense of an integral over some domain determined by relation) is smallest.

In next sections some chosen algorithms are presented.

### 4.3 Relation identification algorithms with inequalities

Let us consider following relational model with inequalities:

\[ R_a(x, y) = \{(x, y) \in X \times Y : f_2(x; a_1) \leq y \leq f_2(x; a_2)\}, \]

where:
\( f_1, f_2 \) – functions known in accuracy to unknown parameters.

In this case we have to identify both functions separately according to our problem statement. Therefore it is enough to present methods for one inequality, because for second one they are analogical but with different sign.

We consider following cases:

- relation with one inequality for SISO system with one unknown parameter;
- relation with one inequality for MISO system with unknown parameters;
- relation with circle inequality for SISO system;
- relation with ellipse inequality for SISO system.

#### 4.3.1 SISO system with one unknown parameter

Let us consider following relational model:

\[ R_a(x, y) = \{(x, y) \in X \times Y : y < a \cdot x\}, \]

where:
\( x \in X \subseteq \mathbb{R}^1; \)
\( y \in Y \subseteq \mathbb{R}^1; \)
\( a \in A \subseteq \mathbb{R}^1. \)

For such problem we can propose following algorithm (based on [5]).
CHAPTER 4. PARAMETER ESTIMATION

Procedure 4.1. Relation identification for SISO system with one unknown parameter

INPUT: $R_N$, input is numeric, output is numeric, $R_a$.

START: $D_a(N) \equiv \emptyset$.

STEP 1
For all examples, $n = 1, 2, \ldots, N$, if $x_n \neq 0$ (identifiability condition), calculate following sets:

$$D_a^{(n)} = \begin{cases} a > \frac{y_n}{x_n}, & \text{if } x_n > 0 \\ a < \frac{y_n}{x_n}, & \text{if } x_n < 0 \end{cases}$$

STEP 2
Calculate $D_a(N)$ as follows:

$$D_a(N) = \bigcap_{n=1}^{N} D_a^{(n)}$$

STEP 3
Determine boundary $\Delta_a(N)$.

OUTPUT: $\Delta_a(N)$.

In general we get

$$D_a(N) = \left( \min_n \left\{ \frac{y_n}{x_n} \right\}, \max_n \left\{ \frac{y_n}{x_n} \right\} \right).$$

Thus the boundary can be easily achieved by taking

$$\Delta_a(N) = \min_n \left\{ \frac{y_n}{x_n} \right\}$$

or

$$\Delta_a(N) = \max_n \left\{ \frac{y_n}{x_n} \right\}$$

or

$$\Delta_a(N) = \frac{1}{2} \left( \min_n \left\{ \frac{y_n}{x_n} \right\} + \max_n \left\{ \frac{y_n}{x_n} \right\} \right)$$

or some other way.
4.3.2 MISO system with unknown parameters

Let us consider following relational model:

\[ R_a(x, y) = \{ (x, y) \in X \times Y : y \leq a^T \cdot x \}, \]

where:
\( x \in X \subseteq \mathbb{R}^S; \)
\( y \in Y \subseteq \mathbb{R}^1; \)
\( a \in A \subseteq \mathbb{R}^K; \)

If \( K \) and \( S \) are not equal, then we have to equalise dimensions of \( x \) and \( a \). Thus:

- If \( K > S \) then we have to extend vector \( x \) by putting 1 in place from \( S + 1 \) to \( K \).
- If \( K < S \) then we have to extend vector \( a \) by putting 1 in place from \( K + 1 \) to \( S \).

For such problem we can use algorithm presented in previous subsection. Then we have

\[ D_a^{(n)} = \{ a : y_n < a^T \cdot x \} \]

and set \( D_a(N) \) is a domain in \( A \) bounded by hyperplanes \( y_n = a^T \cdot x_n \).

To determine boundary we have to use some numerical methods, e.g. linear programming.

However, in particular case with \( K = 2, S = 1 \), it means \( y < a_1 x + a_2 \), we can propose simple numerical method.

Procedure 4.2. MISO system with two unknown parameters

INPUT: \( R_N \), input is numeric, output is numeric, \( R_a \), auxiliary set \( D \).

START: \( D_a(N) \equiv \emptyset, D \equiv \emptyset, i = 1. \)

STEP 1
Take \( i^{th} \) example.
For all \( j > i \) calculate unknown parameters \( a_1, a_2 \) by solving following set of equalities:
\[
\begin{align*}
  y_i &= a_1 x_i + a_2, \\
  y_j &= a_1 x_j + a_2,
\end{align*}
\]
and add them to the set \( D \).

**STEP 2**
If \( i = N - 1 \), then GOTO STEP 3.
Else add to \( D_a(N) \) all parameters’ values from set \( D \) for which all examples are included in \( R_a \), \( D \equiv \emptyset \), \( i = i + 1 \) and GOTO STEP 1.

**STEP 3**
Determine boundary \( \Delta_a(N) \).

**OUTPUT:** \( \Delta_a(N) \).

Here determining boundary is not obvious and could be made randomly by 'blindly' picking one vector of parameters' values or due to some other conditions.

### 4.3.3 Inequality of the circle

Now let us consider an inequality of a circle in SISO system:

\[
R_a(x, y) = \left\{ (x, y) \in X \times Y : (x - x_0)^2 + (y + y_0)^2 < a^2 \right\},
\]
where:

- \( a \) – radius of the circle;
- \((x_0, y_0)\) – center of the circle

*Remark 4.1.* In the case when center of a circle is not \((0, 0)\), it is easy to shift the circle by a circle center to get canonical form of the circle. And the circle center can be calculated as average values of \( x \) and \( y \). Therefore in further considerations we assume that \((x_0, y_0) = (0, 0)\).

In this case the procedure is very simple.

**Procedure 4.3. Circle inequality**

**INPUT:** \( R_N \), input is numeric, output is numeric, \( R_a \).
START: $\mathcal{D}_a(N) \equiv \emptyset$.

STEP 1
Calculate the radius as follows ($d(\cdot, \cdot)$ is a distance measure, e.g. Euclidean metric):

$$a_N = \sqrt{\max_n (x_n^2 + y_n^2)}.$$

STEP 2

$$\mathcal{D}_a(N) = \Delta_a(N) = \{a_N\}.$$

OUTPUT: $\Delta_a(N)$.

For the circle relation we always get only one solution because the circle with such calculated radius includes all examples and has smallest volume.

This simple procedure is very intuitive because the most distance between the most distant example from the center of the circle and the center of the circle determines the radius. For such determined radius all examples are included by the relation.

### 4.3.4 Inequality of the ellipse

Now let us consider an inequality of an ellipse in SISO system:

$$R_a(x, y) = \left\{(x, y) \in X \times Y : \frac{x^2}{a^2} + \frac{y^2}{b^2} < 1 \right\}$$

where:

$a, b$ – radii of an ellipse.

In this case the procedure is a little bit more complicated as it was for the circle. However, idea is similar because we take two the most distant examples from the ellipse center for which parameters's values are calculated.

**Remark 4.2.** Considered model is in the canonical form of an ellipse. Because all bivariate quadratic curves of ellipse type can be transformed into canonical form by shift and rotation \[49\], then we consider only the canonical form.
**Procedure 4.4. Ellipse inequality**

**INPUT:** $R_N$, input is numeric, output is numeric, $R_a$.

**START:** $\mathcal{D}_a(N) \equiv \emptyset$.

**STEP 1**
Choose two examples which are the most distant from the center ($d(\cdot, \cdot)$ is a distance measure, e.g. Euclidean metric),

\[
d((x^{(1)}, y^{(1)}), (0, 0)) < d((x^{(2)}, y^{(2)}), (0, 0)) < \ldots
\]

\[
\ldots < d((x^{(N-1)}, y^{(N-1)}), (0, 0)) < d((x^{(N)}, y^{(N)}), (0, 0)).
\]

**STEP 2**
Calculate parameters $a, b$ from the set of equalities:

\[
\begin{align*}
\left\{ \begin{array}{l}
\frac{(x^{(N-1)})^2}{a^2} + \frac{(y^{(N-1)})^2}{b^2} = 1 \\
\frac{(x^{(N)})^2}{a^2} + \frac{(y^{(N)})^2}{b^2} = 1
\end{array} \right.,
\end{align*}
\]

**STEP 3**

\[
\mathcal{D}_a(N) = \Delta_a(N) = \{a_N\}.
\]

**OUTPUT:** $\Delta_a(N)$.

■

The same as it was in the case of the circle inequality – we get only one solution.
Chapter 5

Relation identification problem for optimal model selection

In this chapter several algorithms for problem of optimal model selection are presented. Moreover, both positive and negative examples with only numeric input and numeric output are considered.

5.1 Problem outline

In chapter 3 a general method for relation identification for both positive and negative examples was formulated. However, for different relation models different algorithms have to be proposed [5, 50]. Nevertheless, very important issue is to formulate special numerical procedure for optimizing criterion model because analytical approach is too complicated or even impossible to apply [5, 50, 55].

5.2 Problem formulation

Problem formulation is the same as it was presented in Problem statement 3.2. Let us briefly remind this problem.

There are given both positive and negative examples, $R_N = R_{N,0} \cup R_{N,1}$, numeric input and output, approximate model with unknown parameters $a \in A$ and some criterion. We are interested in finding such parameters values, $D_a(N)$, for which following implication is fulfilled:

$$a^* \in D_a(N) \Rightarrow Q(a^*) = \min_{a \in A} Q(a).$$
In other words we want to find such relation that covers as many as possible positive examples and includes as small number of negative examples as possible. In ideal situation – relation covers all positive and no negative examples.

The same as it was in previous chapter, we look only for a boundary of \( \mathcal{D}_a(N) \), \( \Delta_a(N) \), which is a singleton, \( \Delta_a(N) = \{ a_N \} \).

In next points some chosen algorithms are presented.

### 5.3 Relation identification algorithms

In our further considerations no more than two classes are taken, it means – observations with \( \lambda = 1 \) and \( \lambda = 0 \). Because for \( L > 2 \) classes it is enough to set all observations with class \( j \) as \( \lambda = 1 \), and for all examples with class \( l \neq j \) as \( \lambda = 0 \) and apply an algorithm. To achieve relations for other classes the same procedure can be used for other \( j \).

In general we consider criterion (2.11). According to it let us take following criterion:

\[
Q(a) = \sum_{n=1}^{N} d[(x_n, y_n), R_a],
\]

where:
\( d(\cdot, \cdot) \) – distance from point to the set.

This distance metric need to be specified. For example it can be zero-one metric and we get criterion (2.12), or distance ‘in direction of \( y \)’ [5]

\[
Q(a) = \sum_{n=1}^{N} \min_{\tilde{y} \in D_y(a, x_n)} d_y(y_n, \tilde{y}),
\]

where:
\( d_y \) – distance ‘in direction of \( y \)’, e.g. \( d(y, \tilde{y}) = ||y - \tilde{y}|| \), or \( d(y, \tilde{y}) = ||y - \tilde{y}||^2 \);
\( \tilde{y} \) – smallest output from relational model;
\( D_y(a, x_n) = \{ y : (x, y) \in R_a \} \).

Then to achieve unknown parameters the criterion has to be minimized. The result depends on observations \( R_N \),

\[
a_N = \Psi(R_N),
\]

where \( \Psi(\cdot) \) is an identification algorithm.
In many cases minimizing the criterion (5.1) or any other criterion is very difficult or even impossible [5]. Therefore some special numerical methods have to be formulated.

In next subsections three numerical algorithms are presented.

5.3.1 Inequality

Let us consider following approximate relation:

\[ R_a(x, y) = \{(x, y) \in X \times Y : y < f(x; a)\} \]

where:
\( x \in X \subseteq \mathbb{R}^S; \)
\( y \in Y \subseteq \mathbb{R}^1; \)
\( a \in A \subseteq \mathbb{R}^K. \)

There are only positive observations and criterion is the same as (5.1). If we take \( d_y[y_n, a \cdot x_n] = |y_n - a \cdot x_n| \), then the criterion for considered relation could be written as follows [5]:

\[ Q(a) = \frac{1}{2} \sum_{n=1}^{N} \left[ |y_n - f(x_n; a)| + y_n - f(x_n; a) \right]. \]

For such problem we can propose following identification algorithm [5].

Procedure 5.1. Optimal model selection for positive examples

**INPUT:** \( R_N \), input is numeric, output is numeric, \( R_a, Q(\cdot) \).

**START:** \( l = 1, \varepsilon. \)

**STEP 1**
Set randomly \( a_0 \), calculate \( Q(a_0) \), and set \( l = l + 1. \)

**STEP 2**
Calculate
\[ a_l = \Psi(a_{l-1}, Q(a_{l-1})). \]
CHAPTER 5. OPTIMAL MODEL SELECTION

STEP 3
If \( |a_t - a_{t-1}| < \varepsilon \), then \( a_N = a_t \) and STOP.
Else GOTO STEP 2.

OUTPUT: \( \Delta_a(N) = \{a_N\} \).

Remark 5.1. The constant \( \varepsilon \) is default, \( \Psi \) is some method of changing unknown parameters, e.g. \( a_t = a_{t-1} \pm \beta, \beta \) – a small value.

5.3.2 Inequality of the circle - two classes

Let us consider an inequality of a circle in SISO system:

\[
R_a(x, y) = \{ (x, y) \in X \times Y : (x - x_0)^2 + (y - y_0)^2 \leq a^2 \}
\]

where:
\( a \) – radius of the circle;
\((x_0, y_0)\) – center of the circle.

Examples are from two classes, \( \lambda_n = 1 \) or \( \lambda_n = 0 \).
In this case we use misclassification criterion (2.10).

Procedure 5.2. Circle inequality with two classes

INPUT: \( R_N \), input is numeric, output is numeric, \( R_a \), \( Q(\cdot) \), metric \( d(\cdot, \cdot) \).

START: \( D_a(N) \equiv \emptyset \).

STEP 1
Calculate the center of the circle:

\[
x_0 = \frac{1}{\text{card}\{R_{N,1}\}} \sum_{i=1}^{\text{card}\{R_{N,1}\}} x_i
\]

\[
y_0 = \frac{1}{\text{card}\{R_{N,1}\}} \sum_{i=1}^{\text{card}\{R_{N,1}\}} y_i
\]
STEP 2
Calculate the set $\tilde{D}_a(N)$:

$$\tilde{D}_a(N) = \{ a : a_n = d((x_n, y_n), (x_0, y_0)), n = 1, 2, \ldots, N \}$$

where $d$ is a metric (e.g. Euclidean metric).

STEP 3
Calculate the set $\tilde{Q}$:

$$\tilde{Q} = \{ Q : \forall a \in \tilde{D}_a(N) \ Q(a) \}$$

STEP 4
Calculate the parameters for which criterion is smallest:

$$a_N \rightarrow Q(a_N) = \min \{ \tilde{Q} \}$$

OUTPUT: $\Delta_a(N) = \{a_N\}$.

After applying this procedure we always get only one vector of unknown parameters.

5.3.3 Inequality of the ellipse - two classes

Let us consider an inequality of an ellipse in SISO system:

$$R_a(x, y) = \{(x, y) \in X \times Y : A \cdot x^2 + 2B \cdot x \cdot y + C \cdot y^2 + D \cdot x + E \cdot y + F \leq 0 \}$$

where:

$a = [A, B, C, D, E, F]$ – parameters.

Examples are from two classes, $\lambda_n = 1$ or $\lambda_n = 0$ and we use misclassification criterion (2.10).

However, it is possible to consider simpler relation by transforming (by shifting and rotating) ellipse inequality into canonical form [49]. Then we can consider only two parameters. Therefore we use following notations:
• \( f \) – shift operation, \( f((x, y), (u, v)) = (x - u, y - v) \);

• \( g \) – rotating operation by angle \( \alpha \), \( g((x, y), \alpha) = (x \cdot \cos(\alpha) - y \cdot \sin(\alpha), x \cdot \sin(\alpha) + y \cdot \cos(\alpha)) \).

Moreover, we need some auxiliary notations:

• \( P = R_{N,1} \) – auxiliary set;

• \( \delta_a(N) \) – set of temporary results which are quadruples \((a, b, \alpha, Q)\), where \( a, b \) – parameters of ellipse in canonical form; \( \alpha \) – angle of rotation for parameters \( a, b \); \( Q \) – value of criterion for parameters \( a, b \).

Procedure 5.3. **Ellipse inequality with two classes**

**INPUT:** \( R_N \), input is numeric, output is numeric, \( R_a, Q(\cdot) \), metric \( d(\cdot, \cdot) \), shift \( f \), rotation \( g \), auxiliary set \( P \), temporary set of results \( \delta_a(N) \).

**START:** \( D_a(N) \equiv \emptyset \).

**STEP 1**

Calculate the center of the ellipse:

\[
x_0 = \frac{1}{\text{card}(R_{N,1})} \sum_{i=1}^{\text{card}(R_{N,1})} x_i
\]

\[
y_0 = \frac{1}{\text{card}(R_{N,1})} \sum_{i=1}^{\text{card}(R_{N,1})} y_i
\]

**STEP 2**

Shift observations:

\[
f(R_{N,0}, (x_0, y_0)) = R'_{N,0},
\]

\[
f(R_{N,1}, (x_0, y_0)) = R'_{N,1},
\]

\[
f(P, (x_0, y_0)) = P'.
\]

**STEP 3**

Find the most distant point from \( O = (0, 0) \) to \( P' \),
\[(x_l, y_l) = \text{arg max}_{(x_k, y_k) \in P'} \{ d((x_k, y_k), O) \}\]

and calculate \(a\):

\[a = d((x_l, y_l), O)\]

**STEP 4**

Calculate \(\alpha\):

- if \((x_l > 0 \text{ and } y_l > 0)\), then \(\alpha = \arcsin \frac{y_l}{\sqrt{x_l^2 + y_l^2}}\);
- if \((x_l < 0 \text{ and } y_l > 0)\), then \(\alpha = 180 - \arcsin \frac{y_l}{\sqrt{x_l^2 + y_l^2}}\);
- if \((x_l > 0 \text{ and } y_l < 0)\), then \(\alpha = \arcsin \frac{y_l}{\sqrt{x_l^2 + y_l^2}}\);
- if \((x_l < 0 \text{ and } y_l < 0)\), then \(\alpha = -\arcsin \frac{y_l}{\sqrt{x_l^2 + y_l^2}}\).

Remove point \((x_l, y_l)\) from \(P'\),

\[P' = P' \setminus \{(x_l, y_l)\}\]

Rotate observations by angle \(\alpha\),

\[g(R_{N,0}', \alpha) = R_{N,0}''\]
\[g(R_{N,1}', \alpha) = R_{N,1}''\]
\[g(P', \alpha) = P''\]

**STEP 5**

Find \(b\):

\[\forall (x_p, y_p) \in P'' \text{ if } y_p \neq 0, \text{ then } b_p = \sqrt{\frac{a^2 \cdot y_p^2}{a^2 - x_p^2}}\]

Calculate criterion of misclassification:

\[\forall b_p \ Q(a, b_p)\]

and choose such \(b\) for which this criterion is minimal:

\[Q(a, b) = \min_{b_p} \{ Q(a, b_p) \} \]
STEP 6
Add calculate quadruple to the set $\delta_a(N)$.
If $\text{card}\{P'\} = 1$, then GOTO STEP 7.
Else – GOTO 3.

STEP 7
Find smallest criterion value among quadruples in $\delta_a(N)$:

$$Q_{\text{min}} \rightarrow \delta_{a,\text{min}}(N) = \min_{Q} \{\delta_a(N)\}.$$ 

Add to the boundary $\Delta_a(N)$ all quadruples for which $Q_{\text{min}},$

$$\Delta_a(N) = \{(a, b, \alpha) : Q(a, b) = Q_{\text{min}}\}$$

and STOP.

OUTPUT: $\Delta_a(N)$.

It is very easy to transform canonical form to the ellipse in form of bivariate quadratic curve if $(x_0, y_0)$ and $\alpha$ are known.

Besides, the boundary is not a singleton. Therefore we can choose randomly one vector of parameters or propose some other criterion (e.g. volume of the relation) to choose one vector.
Chapter 6

Knowledge representation with additional uncertainty descriptions

In this chapter some algorithms for relation identification with additional uncertainty descriptions. Because there is a vast of different certainty descriptions and for each of them different approaches are needed, therefore the idea of relation identification is considered only for random variables. However, we present problem statement in a general form, which is valid for different uncertainty descriptions.

6.1 Introduction

In some situations of knowledge discovery it is possible to use additional knowledge which is formulated by uncertainty description. There are several uncertainty descriptions:

- random variables [1, 9, 10, 22, 23, 46];
- fuzzy variables and sets [9, 10, 16];
- uncertain variables [9, 10, 11, 59];
- rough sets [40, 41].

Prof. Bubnicki proposed in his works some general description of uncertainty called soft variables (e.g. [10, 11]). This variable is defined by an evaluating function which fulfills several properties.
**Definition 6.1. (Soft variable and the pair of soft variables)** The soft variable

\[ \hat{x} = < X, g(x) > \]

is defined by the set of values \( X \) and a bounded evaluating function \( q : X \rightarrow \mathbb{R}^+ \), satisfying following condition:

\[
\int_X x \cdot g(x) < \infty
\]

for the continuous case and

\[
\sum_{i=1}^{\infty} x_i \cdot g(x_i) < \infty
\]

for the discrete case.

Let us consider two soft variables \( \hat{x} = < X, g(x) > \), \( \hat{y} = < Y, g(y) > \) and the variable \((\hat{x}, \hat{y})\) described by \( q_{xy}(x, y) : X \times Y \rightarrow \mathbb{R}^+ \). The conditional evaluating function is denoted by \( g_y(y|x) \) which is the evaluating function of \( \hat{y} \) for the given value \( (\hat{x} = x) \).

The pair \((\hat{x}, \hat{y})\) is defined by \( g_{xy}(x, y) \) and two operations:

\[
g_{xy}(x, y) = O_1[g_x(x), g_y(y|x)],
\]

\[
g_x(x) = O_2[g_x(x), g_{xy}(x, y)],
\]

e.g.

\[
O_1 : D_{gx} \times D_{gy} \rightarrow D_{g,xy}, \quad O_2 : D_{g,xy} \rightarrow D_{gx}
\]

where \( D_{gx}, D_{gy} \) and \( D_{g,xy} \) are sets of the functions \( g_x(x), g_y(y|x) \) and \( g_{xy}(x, y) \), respectively.

**Definition 6.2. (Mean value of soft variable)** Let us consider a soft variable

\[ \hat{x} = < X, g(x) > . \]

Then the mean value can be defined as follows:

\[
M(\hat{x}) = \frac{\int_X x \cdot g(x)dx}{\int_X g(x)dx}
\]
in continuous case $X$, and

$$M(\hat{x}) = \frac{\sum_{k=1}^{K} x \cdot g(x)}{\sum_{k=1}^{K} g(x)}$$

in discrete case, $X = \{x_1, x_2, \ldots, x_K\}$.

The evaluating function may have different practical interpretations [10, 11, 12]. In the random case, a soft variable is a random variable described by the probability density function. In the case of the fuzzy description, a soft variable is a fuzzy variable described by the membership function. In the case of an uncertain variable, evaluating function is the certainty distribution. In the case of rough set, the evaluating function is a rough-membership function. Thus in general it could be said that evaluating function describes an evaluation of the set of possible values $X$, characterizing for every value $x$ its significance (importance or weight).

The problem statement could be based on the idea of the soft variable. However, we present algorithms for case of random variables.

### 6.1.1 Problem outline

For the relation identification problem with the uncertain description we can distinguish two kinds of problems, the same as it was made in previous chapters of this works,

2. Optimal model selection [5, 13, 50, 55, 58, 60, 61].

In the first case we assume that there is given description of the system known in accuracy to unknown parameters, it means:

$$R_a(x, y) = \{(x, y) \in X \times Y : w[F(x, y; a)] = 1\},$$  \hspace{1cm} (6.1)

$$F(x, y; a),$$  \hspace{1cm} (6.2)

$$g(x, y; a)$$  \hspace{1cm} (6.3)

where:

$R_a$ – relational model;
$F(x, y; a)$ – complex relation description known in accuracy to unknown parameters;
$g(x, y; a)$ – evaluating function known in accuracy to unknown parameters;
$a \in A \subseteq \mathbb{R}^K$ – unknown parameters.

Besides, observations are only positive, $\lambda_n = 1$ for $n = 1, 2, \ldots, N$.

In the second case we assume that observations could belong to the relation, $\lambda_n = 1$, or not – $\lambda_n = 0$. Moreover, we have an approximation of the relation in form of (6.1) given by an expert. Furthermore, there exists an evaluating function $g(x, y)$ for soft variables $(\tilde{x}, \tilde{y})$ and $g(x, y; a)$ defined on the $R_a$ is given. Moreover, we do not know exact value of $\lambda$ and we say that it is a soft variable, $\tilde{\lambda}$. Therefore we can determine belonging in condition of observation:

$$g(\tilde{\lambda} = 1|\tilde{x} = x, \tilde{y} = y) = g_1(x, y).$$  \hspace{1cm} (6.4)

To calculate unknown parameters some criterion has to be proposed.

### 6.1.2 Problem formulation

Let us formulate two problems mentioned in previous subsection.

**Problem Statement 6.1. Parameter estimation**

**GIVEN:**

- Relational model $R_a$.
- Evaluating function $g(x, y; a)$.
- Examples $R_{N,1}$.

**FIND:**

- Unknown parameters $a$ for which relational model covers as much examples as possible due to the evaluating function.
Problem Statement 6.2. **Optimal model selection**

**GIVEN:**
- Relational model $R_a$.
- Evaluating function $g(x, y; a)$.
- Examples $R_N$.
- Evaluating function $g_1(x, y)$.
- Criterion $Q(\cdot)$.

**FIND:**
- Unknown parameters $a^*$ for which criterion is minimal:
  
  $a^* \rightarrow Q(a^*) = \min_{a \in A} \{Q(a)\}$.

### 6.2 Relation identification algorithms

In next points we present two well known methods for parameter estimation, which are extensions of methods for functional description [3], and one general consideration about model selection problem. Both methods are formulated for random variables.

#### 6.2.1 Parameter estimation approach

**Maximum likelihood method**

Let us assume that input and output are random variables, $(x, y)$, described by a probability density function known in accuracy to unknown parameters, $g(x, y; a)$. Moreover, observations are independent realizations of $(x, y)$.

We can formulate a likelihood function:

$$W(a; R_N) = \prod_{n=1}^{N} g(x_n, y_n; a).$$  \hfill (6.5)
The estimation of parameters are achieved by maximizing (6.6) due to $a$, it means:

$$a_N = \Psi(R_N) \longrightarrow W(a_N; R_N) = \max_{a \in A} \left\{ W(a; R_N) \right\}. \quad (6.6)$$

where:

$\Psi$ – algorithm of estimation.

**Example 6.1.** Let us consider probability density function in following form:

$$g(x,y; a) = \begin{cases} \left( \int_{R_a} dx dy \right)^{-1}, & \text{for } (x, y) \in R_a \\ 0, & \text{otherwise} \end{cases},$$

Then the likelihood function is as follows:

$$W(a, R_N) = \begin{cases} \left( \int_{R_a} dx dy \right)^{-N}, & \text{if } \forall_{n=1,2,...,N}(x_n, y_n) \in R_a \\ 0, & \text{otherwise} \end{cases},$$

Then the determination of the estimator is equivalent to determination of such parameters $a$ for which all points belong to $R_a$ and the "volume" of $R_a$ is minimal, it means:

$$a_N \longrightarrow \min_{a \in A} \left\{ \int_{R_a} dx dy \right\}$$

with conditions $w[F(x_n, y_n; a)] = 1, n = 1, 2, \ldots, N$.

At the end it is interesting to mention that if we assume the same probability density function as in the Example 6.1 and we take the circle or ellipse inequality, then methods presented in chapter 4 of this work are algorithms fulfilling the condition of minimal "volume" and including all observations.

This remark means that intuitive approach presented in chapter 4 is equivalent to the conclusions of the well known method. Moreover, it shows that conclusions drawn from the maximum likelihood method are very useful for formulating identification algorithms for different forms of relational models [58].

**Bayes’ Method**

Now let us consider very briefly other very well known method - *Bayes’ method*. If we assume that unknown parameters $a$ are realizations of a random variable $a$ with known probability density function $g_a(a)$, then we can use Bayes’ method. It is a known fact ([3, 58]) that the estimation problem is equivalent to the minimizing of a conditional risk. Thus we have:
\[ a_N = \Psi(R_N) \rightarrow r(a_N; R_N) = \max_{\bar{a} \in A} \{ r(\bar{a}; R_N) \}. \] (6.7)

where:
\( r(\cdot, \cdot) \) – conditional risk,
\[ r(\bar{a}; R_N) = \int_A L(a, \bar{a}) \cdot g_a(a) \prod_{n=1}^N g(x_n, y_n | a) da \]
\( L(\cdot, \cdot) \) – loss function.

### 6.2.2 Optimal model selection approach

In the optimal model selection we can distinguish two problems:

1. Complete knowledge about uncertainty.
2. Incomplete knowledge about uncertainty.

In the first case we assume that \( g(x, y), g_1(x, y) \) are known and \( R_a, g_a(x, y) \) are given by an expert.

In the second case we assume that only \( R_a, g_a(x, y) \) and \( R_N \) are given.

For both cases we formulate only general idea of the relation identification. Based on our considerations specific algorithms can be proposed.

#### Complete knowledge about uncertainty

Having \( g(x, y), g_1(x, y), R_a \) and \( g_a(x, y) \) we can proposed following probability criterion:

\[ Q(a) = \mathbb{E} q \left( \lambda, w \left[ F(x, y; a) \right] \right) \] (6.8)

where:
\( \mathbb{E} \) – expected value (see definition 6.2);
\( q(\cdot, \cdot) \) – dissimilarity measure (2.9).

Calculating the (6.8) we get

\[ Q(a) = \int_{X \times Y} \left[ q(\lambda = 1, w[F(x, y; a)]) \cdot g_1(x, y) + \ight. \\
+ q(\lambda = 0, w[F(x, y; a)]) \cdot g_0(x, y) \right] \cdot g(x, y) dxdy \]
where:
\[ g_0(x, y) = 1 - g_1(x, y). \]

The optimal values of parameters of the model can be determined by minimizing criterion \(6.8\),
\[ a^* \rightarrow Q(a^*) = \min_{a \in A} \{ Q(a) \}. \]

**Incomplete knowledge about uncertainty**

Having \( R_a, g_a(x, y) \) and \( R_N \) we can propose the following empirical probability criterion:

\[
Q(a) = \frac{1}{N} \left[ \sum_{n \in E_{N,1}} \varrho[(x_n, y_n), R_a] \cdot \frac{N_{n1}}{N_n} + \sum_{n \in E_{N,0}} \varrho[(x_n, y_n), \neg R_a] \cdot \frac{N_{n0}}{N_n} \right] \tag{6.9}
\]

where:
- \( \varrho \) – metric;
- \( N_n \) – number of repeats of observation \((x, y)\);
- \( N_{n1} \) – number of repeats of observation \((x, y)\) with \( \lambda_n = 1 \);
- \( N_{n0} \) – number of repeats of observation \((x, y)\) with \( \lambda_n = 0 \).

The optimal values of parameters of the model we get by minimizing criterion \(6.9\),
\[ a^* \rightarrow Q(a^*) = \min_{a \in A} \{ Q(a) \}. \]
Chapter 7

Applications of relation identification methods

In this chapter three applications of relation identification methods are presented. Examples of applications represent problems formulated in the work, it means:

- Parameter estimation.
- Optimal model selection.
- Modified coverage strategy with optimal model selection.

Case 1 is an application of the procedure 3.1 with no nominal attributes in economy (simulation).

Case 2 is an application of the procedure 3.3 and 3.2 with no nominal attributes in biomedical case (real-life dataset).

Case 3 is an application of the procedure 3.3 and 3.2 with no nominal attributes in the well-known benchmark dataset.

Moreover, each application shows that usage of relational knowledge representation is adequate and can model reality so that it could be very useful in decision making or analysis problem.

All relation identification methods were implemented in Matlab®. To compare them with other data mining methods WEKA free software [66] was used.
7.1 Relational knowledge representation in economy

7.1.1 Problem description

First example of RKR’s application is used in the economy. Let us consider a problem of supply management \[37\]. The main goal in this problem is to gather supplies in the warehouse so that there is always some reserve but cost of storage is minimal. In other words, there is a need to calculate some optimal level of supplies with minimal cost of storage.

After some analysis and empirical studies following equation of costs was formulated \[37\]:

\[ C = 0.5k \cdot D + (r + s \cdot D) \cdot \frac{Q}{D}, \tag{7.1} \]

where:

- \( C \) – total average cost of storage in the year (output);
- \( D \) – total average number of delivered units (level of supplies) in the year (input);
- \( Q \) – average number of sold units in the year;
- \( k \) – change of storage cost of one unit during the year;
- \( r \) – cost of one delivery;
- \( s \) – change of cost per one unit.

To understand this equation let us consider its separate parts \[37\]:

- \( 0.5k \cdot D \) – is average cost of storing during a year;
- \( r + s \cdot D \) – is average cost per one delivery;
- \( \frac{Q}{D} \) – is average number of deliveries.

If it is possible to determine unknown parameters \( Q, k, r, s \), then it is a simple problem of optimizing the equation (7.1) and we get:

\[ D^* = \sqrt{\frac{2 \cdot r \cdot Q}{k}}. \]

And depending on values of parameters we get different optimal decisions. However, we always get one optimal decision.

In real-life situations using such equation could be applied only if we know all parameters. Therefore identification methods should be applied \[3, 33, 55\].

Nevertheless, this equation gives a solution in average sense. It means that even if a decision-maker, e.g. warehouse manager, has some additional knowledge about
the warehouse, the functional knowledge representation (7.1) does not give him an opportunity to use this knowledge. Therefore, some more general description should be used.

### 7.1.2 Relational model

Now let us assume that the warehouse was observed during several years and it is possible to determine $k$, $r$, and $s$ as a constant values. However, $Q$ got different values during those several years. In such case we propose following relation:

$$R_{a}(C, D) = \begin{cases} (C, D) \in \mathbb{R}^2 : 0.5k \cdot D + (r + s \cdot D) \cdot \frac{a_1}{D} < C < 0.5k \cdot D + (r + s \cdot D) \cdot \frac{a_2}{D} \end{cases}$$

(7.2)

where:

$a_1, a_2 \in \mathbb{R}^1$ - unknown parameters.

Then we can use the procedure 3.1 with $D$ as an input and $C$ as an output and the idea of the procedure 4.1. Thus unknown parameters can be determined as follows:

$$a_1 = \min_{n=1,2,...,N} \left\{ \frac{C_n \cdot D_n - 0.5k \cdot D_n}{r + s \cdot D_n} \right\}$$

and

$$a_2 = \max_{n=1,2,...,N} \left\{ \frac{C_n \cdot D_n - 0.5k \cdot D_n}{r + s \cdot D_n} \right\}.$$

To illustrate the problem (simulation) let us take following values:

- $k = 8$;
- $r = 20$;
- $s = 3$;
- $Q$ was randomly drawn using uniform distribution from $[190; 210]$;
- average number of units delivered in the year, $D$, was generated using normal distribution with mean value 30 and variance 7.5 (to avoid real values *ceil function* was used) and total number of observation was 30 (which can be understood that observations were collected during 3 years from 10 warehouses).

Generated observations are presented in the figure [7.1]. The background knowledge is following: $D \in [10, 60]$.

After applying mentioned procedures and parameters' determination we get following values:
Figure 7.1: Dataset of storage management.

- \( a_1 = 190; \)
- \( a_2 = 210; \)
- \( a_{\text{mean}} = 200.5 \) (calculated as a mean value of all values of \( a_1 \) and \( a_2 \)).

The relation is following:

\[
R_a(C, D) = \{ (C, D) \in \mathbb{R}^2 : (D \in [10, 60]) \land \\
\quad 0.5k \cdot D + (r + s \cdot D) \cdot \frac{190}{D} < C < 0.5k \cdot D + (r + s \cdot D) \cdot \frac{210}{D} \}.
\]

In the figure 7.2 the relation is presented and function with \( Q = 200 \) as a comparison.

### 7.1.3 Conclusions

It is remarkable to stress the usefulness of usage of relational knowledge representation. Using the functional knowledge representation the warehouse manager can make only decisions which are optimal due to the model (7.1). However, using any additional knowledge or intuition of the manager cannot be used or there are no premises of using it. Therefore some other description need to be formulated.
As it is showed using the simulation, relational description can be very useful. Using RKR the manager can analyse current situation on the market and can predict which decision (value of $D$) should he make and what costs $C$ he can expect between minimal and maximal value. The final cost depends on some unknown factors.

Those factors could be presented on the example with parameters used in the simulation ($k = 8$, $r = 20$, $s = 3$, $Q \in [190, 210]$). For the functional description the optimal $D^*$ is 32 and $C^*$ equals 853. Now let us assume that warehouse manager signed a contract with a new client for a new year. Then he can make order for more supplies but with restriction to get costs no bigger than 853. Using relational description he can calculate costs in other way. If he assumes that the time of storage will be short because of the new client (optimistic case), then he can calculate costs by taking lower limit and order between 32 and 53 units (then costs will be between 816.75 for 32 and 853 for 53). However, final decision should be made by taking into consideration the possible risk because in worst case the cost will be between 889.2 for 32 (the lowest risk) and 921.25 for 53 (the highest risk).

This simple example shows that the manager has more possible decisions to make and due to his additional knowledge he can get more profits. Moreover, he can be prepared for different scenarios and calculate final profit for best, worst or
other cases.

Besides, the relation is a generalization of the function. Firstly, values of the function belongs to the domain determined by the relation. Secondly, if we take the same form of the function as an upper and lower limitation of the relation, then calculating mean value of the unknown parameters gives as the parameter of the function (see results in subsection 7.1.2).

At the end it is worth to say that we can achieve the same or similar results if we apply methods with additional uncertainty knowledge. Results depend on the available uncertainty description.

### 7.2 Relational knowledge representation in biomedical case

#### 7.2.1 Problem description

Second example is connected with biomedical case. We consider problem of cerebral aneurysms [24]. The aneurysm "is an abnormal localized dilation of any vessel" [63] and "it most commonly occurs in arteries that supply blood to the brain" [63]. Besides, this illness is very dangerous because if it ruptures it causes serious brain injuries or even death.

Moreover, it is estimated that cerebral aneurysms occur in up to 5−6% of general human population [24, 63]. Therefore it is very important to conduct continuous researches about this illness. One way of trying to understand unstable behaviour of cerebral aneurysms is mathematical modelling.

Up now there were several models proposed (e.g. [24]) and most of them are based on differential equations or calculating some function describing risk of rupture. We propose other approach of cerebral aneurysm’s modelling using relational description.

Due to the lack of data we describe a patient by:

- **Input:**
  1. $x^{(1)}$: age of a patient = numeric;
  2. $x^{(2)}$: size of an aneurysm (diameter) = numeric.
CHAPTER 7. APPLICATIONS OF RELATION IDENTIFICATION METHODS

Output:

1. $y$: ruptured = {YES, NO}

We assume that age is the first input $x^{(1)} \in [0, 90]$ and size is the second input $x^{(2)} \in [0, 30]$ (background knowledge $B = \{(x, y) \in X \times Y : x^{(1)} \in [0, 90] \land x^{(2)} \in [0, 30]\}$), and it can be determined if an aneurysm is ruptured ($\lambda_n = 1$, class ruptured = "YES"), or not ($\lambda_n = 0$, class ruptured = "NO"). And having such relational object we are interested in finding description of higher risk of rupture and lower risk of rupture. Then we can use such extracted knowledge to classify patient to the group of higher risk (and e.g. put him/her on the list to make an operation) or lower risk (and put him/her on the list to treat with medicines). Such decision can be used in computer decision support system in helping making diagnosis by a doctor [62].

The dataset was obtained from literature: [35], [36], [45], [64]. There are 99 instances, 21 from class $\lambda = 1$ and 78 from class $\lambda = 0$. Dataset is presented in the figure 7.3.

![Dataset for biomedical problem](image_url)

Figure 7.3: Dataset for biomedical problem.
7.2.2 Relational model

We are interested in finding such relation for which as many as possible rupture aneurysms and as small as possible non-ruptured aneurysms are included. After analysing dataset (there are more instances for $\lambda = 0$ and instances $\lambda = 0$ are distributed more regularly) let us assume that:

- We use modified coverage knowledge discovery method (procedure 3.3).
- We calculate linear inequality $(x^{(2)} \leq a \cdot x^{(1)} + b)$ with attributes age and size for instances $\lambda = 0$ and complement of that inequality for $\lambda = 1$.
- We use modified procedure 4.2 (using criterion instead of including all observations) with misclassification criterion.

7.2.3 Results

After applying methods mentioned before, we get two following relations:

1. Relation 1:
   - Relation for $\lambda = 0$:
     \[ \tilde{R}_a(x^{(1)}, x^{(2)}) = \{ (x, y) \in X_1 \times X_2 : x^{(2)} \leq 0.1910 \cdot x^{(1)} + 3.0337 \}; \]
   - minimal criterion value:
     \[ Q^* = 16; \]
   - number of examples with $\lambda = 1$ included is 8.

2. Relation 2:
   - Relation for $\lambda = 0$:
     \[ \tilde{R}_a(x^{(1)}, x^{(2)}) = \{ (x, y) \in X_1 \times X_2 : x^{(2)} \leq 0.1667 \cdot x^{(1)} + 3.6667 \}; \]
   - minimal criterion value:
     \[ Q^* = 16; \]
   - number of examples with $\lambda = 1$ included is 7.
Because, from practical point of view, we should choose only one set of unknown parameters, thus some additional criterion could be formulated. Let say that if there are more than one vector of unknown parameters for which criterion is the same minimal value, then choose this vector that relation includes less negative examples.

In our case we choose relation 2 because it includes 7 examples with $\lambda = 1$ which is less than for relation 2.

Let us denote by $\tilde{R}_a^{(0)}$ relation $\tilde{R}_a$ and by $\tilde{R}_a^{(1)}$ – complement of $\tilde{R}_a^{(0)}$, $\tilde{R}_a^{(1)} = \neg\tilde{R}_a^{(0)}$.

Then we get two partial descriptions:

$$\tilde{F}_1 : \tilde{R}_a^{(1)}(x^{(1)}, x^{(2)}) \Rightarrow \text{(ruptured = "YES");}$$

$$\tilde{F}_0 : \tilde{R}_a^{(0)}(x^{(1)}, x^{(2)}) \Rightarrow \text{(ruptured = "NO");}$$

and final relation:

$$R_a(x, y) = \{ (x, y) \in X \times Y : \tilde{F}_0 \land \tilde{F}_1 \land B \}.$$

Two obtained relations, Relation 1 and Relation 2, are presented in the figure 7.4.

### 7.2.4 Comparison

Achieved relational object description can be treated as classification rules. Therefore it is easy to make a comparison between relational approach with well-known methods. WEKA software was used with following methods (more detailed description can be found in [66]):

- JRip, propositional rule learner Repeated Incremental Pruning to Produce Error Reduction (RIPPER);
- NNge, Nearest neighbor like algorithm using non-nested generalized exemplars;
- OneR, 1R classifier;
- PART, PART decision list;
- Ridor, RIpple-DOwn Rule learner;
- Naive Bayes;
- Multilayer Perceptron;
Figure 7.4: Partial relations (dark gray – higher risk of rupture, lighter gray – lower risk of rupture).
• J48 (tree algorithm).

All methods were applied with fixed values before classification. Four criteria to compare methods were used [65]:

• classification accuracy;

• number of instances from class YES classified as YES (true positive) and number of instances from class YES classified as NO (false negative) – true positive rate;

• number of instances from class NO classified as NO (true negative) and number of instances from class NO classified as YES (false positive) – false positive rate;

• Kappa statistics – (also known as Cohen’s kappa coefficient) it is an index which compares the agreement against that which might be expected by chance; possible values range are $[-1, 1]$, where $+1$ is perfect agreement, $0$ is no agreement above that expected by chance, and $-1$ is complete disagreement; it can be calculated as follows:

$$\kappa = \frac{\text{Observed agreement} - \text{Chance agreement}}{1 - \text{Chance agreement}}.$$

The comparison with the relational approach (Relation 1 and Relation 2 in the figure) is shown in the figure [7.5].

### 7.2.5 Conclusions

After comparison it is easy to notice that relational descriptions gives the best results. They have highest classification accuracy and have significantly higher true positive rate (for Relation 2 2/3 of instances from class $\lambda = 1$ were classified correctly; it was marked by a line in the figure [7.5b]). On the other hand the false positive rate is average in comparison to other methods but it is still high (around 90%).

From the medical point of view the most important is true positive rate because treating patient with no rupture ($\lambda = 0$) as with rupture ($\lambda = 1$) makes no harm to anybody. However, from the economical point of view (e.g. if patient is classified to be operated, which implies additional costs of treatment) false positive rate plays more important role. Nevertheless, in all three criteria relational description had very good, and even the best, performance in comparison to other DM methods.
CHAPTER 7. APPLICATIONS OF RELATION IDENTIFICATION METHODS

(a) Accuracy

(b) True positive rate

(c) False positive rate

(d) Kappa statistics

Figure 7.5: Comparison of methods.
Moreover, the kappa statistics’ value for relational descriptions are highest among other methods. Besides, relational description marked as Relation 2 has highest value and it is also a hint to choose this relation as a final description.

At the end it is good to stress that collected dataset (99 instances) cannot be compared with huge datasets (thousands of examples). However, in biomedical applications such amount is acceptable for making serious considerations. Moreover, obtained results could be already used by doctors in real decision support systems due to its accuracy (see figure 7.5a).

7.3 Relational knowledge representation in benchmark dataset

7.3.1 Problem description

Last example is connected with well-known weather dataset. In the literature it is one of the widely used benchmark (see e.g. [48] page 104, [65] page 12). This dataset is very simple and contains only 14 instances. However, because it is the benchmark problem in data mining, it is very convenient to compare new approaches with it. Besides, this small amount of examples causes that any analysis could be made even ”on the paper”, by hand, and any conclusions could be easily checked and understood.

This dataset consists of following attributes:

- Input:
  1. \(x^{(1)}\): outlook = \{ sunny, overcast, rainy \};
  2. \(x^{(2)}\): temperature = numeric;
  3. \(x^{(3)}\): humidity = numeric;
  4. \(x^{(4)}\): windy = \{ FALSE, TRUE \}.

- Output:
  1. \(y\): play = \{ YES, NO \}.

The background knowledge is following:
\[
\mathcal{B} = \{(x, y) \in X \times Y : x^{(2)} \in [0, 100] \land x^{(3)} \in [0, 100]\}.
\]
7.3.2 Relational model

We are interested in finding such relation which describes as good as possible the reality. It means what conditions should be fulfilled to play \((\text{play} = \text{"YES"})\), and what – not \((\text{play} = \text{"NO"})\). After analysing dataset let us assume that:

- We use modified coverage knowledge discovery method (procedure 3.3).
- We calculate ellipse inequality with attributes \textit{temperature} and \textit{humidity} for instances \(\lambda = 1\) and complement of that ellipse for \(\lambda = 0\).
- We use procedure 5.3 (calculating unknown parameters for inequality of ellipse) with \textit{misclassification criterion}.

7.3.3 Results

After applying method for ellipse inequality we get only one following ellipse (see figure 7.6a):

- partial relation in canonical form for \(\lambda = 1\):
  \[
  \tilde{R}_a(x^{(2)}, x^{(3)}) = \left\{ (x^{(2)}, x^{(3)}) \in X_2 \times X_3 : \frac{(x^{(2)})^2}{12.1432^2} + \frac{(x^{(3)})^2}{10.6226^2} \leq 1 \right\};
  \]
- center of the ellipse:
  \[(x_0^{(2)}, x_0^{(3)}) = (73; 79.1111);\]
- angle of rotation:
  \[\alpha = 34.5625;\]
- minimal criterion value:
  \[Q^* = 2.\]

Let us denote by \(\tilde{R}_a^{(1)}\) shifted and rotated relation \(\tilde{R}_a\) and by \(\tilde{R}_a^{(0)} – \text{complement of } \tilde{R}_a^{(1)}\), \(\tilde{R}_a^{(1)} \cap \tilde{R}_a^{(0)} = -\tilde{R}_a^{(1)}\).

After applying modified coverage method we get following partial descriptions (see figure 7.6b):

\[
\tilde{F}_1 : \tilde{R}_a^{(1)}(x^{(2)}, x^{(3)}) \lor (\text{outlook} = \text{"overcast"}) \Rightarrow \text{play} = \text{"YES"};
\]
\[
\tilde{F}_0 : \tilde{R}_a^{(0)}(x^{(2)}, x^{(3)}) \Rightarrow \text{play} = \text{"NO"};
\]
and then criterion value is:

\[ Q^* = 1. \]

Final relation is as follows:

\[ R_a(x, y) = \{(x, y) \in X \times Y : \tilde{F}_0 \land \tilde{F}_1 \land B\}. \]

Figure 7.6: Two steps in partial relation.
7.3.4 Comparison

The proposed relational description was compared with well-known DM methods. There were used the same methods and criteria as presented in subsection 7.2.4 of this work. Results of those comparison are shown in the figure [7.7].

7.3.5 Conclusions

After comparison it is obvious that relational model gives the best results (the same as J48 algorithm). It has highest classification accuracy (see [7.7a]) and highest true positive rate (around 93%, see [7.7b]) and false positive rate (100%, see [7.7c]). Due to inductive reasoning paradigm we can say that almost ideal description was found.

In this case, in contradiction to biomedical problem, both true and negative positive rates are important. Because players want rather to avoid bad weather (e.g. avoid possible illnesses after playing in unpleasant temperature and humidity) than go playing in most situations (opposite situation than in the biomedical case where there is better to classify patient with lower risk to higher; here it is better to stay at home than go playing).

Moreover, the kappa statistics’ value is highest for relational description among other methods and equal with J48 algorithm. Furthermore, kappa value equals 0.85 which implies that relational approach is a very good description for considered problem.

Of course this simple case gives only an opportunity to compare relational description with other methods and make no other conclusions for creating decision support system, or expert system. Nevertheless, results obtained after applying relational approach to this benchmark dataset show that it gives very good description. And due to its possible application in expert system with logic-algebraic method as a reasoning method represents very interesting example of relational methods’ application.
CHAPTER 7. APPLICATIONS OF RELATION IDENTIFICATION METHODS

(a) Accuracy

(b) True positive rate

(c) False positive rate

(d) Kappa statistics

Figure 7.7: Comparison of methods.
Chapter 8

Conclusions

8.1 Final remarks

In this work a coherent survey of problems connected with relational knowledge representation and methods for achieving relational knowledge representation were presented. Moreover, usage of RKR was shown in economic and biomedical applications and was applied to the benchmark dataset. In the main part of this work crucial definitions were presented, like relational system, relational model, dataset. Furthermore, general problem statement and general method of knowledge discovery were formulated. Next three main algorithms were proposed:

- relation identification method with parameter estimation;
- relation identification method with optimal model selection;
- modified coverage knowledge discovery method.

Then specific algorithms for parameter estimation and optimal model selection were proposed. Methods for chosen relational models showed the main idea of creating other algorithms. Some of presented methods were taken from the literature (procedure 4.1, 5.1) and are formulated by author of this work (procedure 4.2, 4.3, 4.4, 5.3, 5.4). Moreover, in case of different observations’ types (only positive, or positive and negative examples) parameter estimation or optimal model selection methods could be used.

Next a problem with additional uncertain knowledge was formulated. General considerations were made for generalized uncertainty description, so called soft variables. However, relation identification methods were formulated for random variables.
At the end three applications of presented approach were shown. In each case different relational models were used. Obtained results showed usefulness of RKR and purposefulness of further consideration and research under developing new methods for relation identification and knowledge extracting. Besides, each RKR achieved in the examples could be already used in creating expert systems with logic-algebraic method as a reasoning method.

Some remark should be stated about evaluating and comparing two RKR. It could be made due to chosen optimal criterion. However, very often some other criteria should be taken and it could be made the same as in the traditional algorithms’ analysis. Nevertheless, sometimes the comparison cannot be made. Especially if relational description is compared with functional description. For example in biomedical application using functional description can be used in calculating some trend of aneurysm’s growing, but it cannot be taken to create logical description with classes.

At the end it should be answered if aims of the thesis were achieved. In other words – if research questions were answered. Thus:

- Research question Q1 – answer to this question could be found in point 2.1 of this work where relational knowledge representation was formulated.

- Research question Q2.1 – answer to this question could be found in points 2.1, 2.2, 2.3 of this work.

- Research question Q2.2 – answer to this question could be found in points 2.9, 3.2.1, 3.2.2, 3.2.3 of this work where general and more detailed problems were formulated.

- Research question Q2.3 – answer to this question could be found in points 2.8, 2.9, 3.2.1, 3.2.2, 3.2.3 of this work where general and more detailed methods were formulated.

- Research question Q2.4 – answer to this question could be found in chapters 4, 5 and 6 of this work.

- Research question Q2.5 – answer to this question could be found in chapters 3, 4, 5, and 6 of this work where problems of parameter estimation, optimal model selection, relation discovery and knowledge discovery with uncertainty description were presented.
8.2 Future work

The main goal of this work was a coherent survey about relational knowledge representations and relation identification methods. In other words this work was meant as a basis for further, more complicated considerations. And formulating new knowledge representation, relational knowledge representation, create a lot of possibilities in describing reality. Furthermore, having logic-algebraic method gives great opportunity to use relational description in practice. Besides, automated relational knowledge acquisition opens "door" to very interesting and useful applications.

Therefore it seems that following topics should be considered in the future:

1. Aspect of experiment for relational system was omitted in this work. In the literature there were some tries to carry out active experiment for relational system (e.g. [27]) but it cannot be treated as a complete study.

2. Very interesting from practical point of view would be creation of expert system with one part with relation identification methods for different relational models and second part with logic-algebraic method as a reasoning method. Then such expert system could be applied in e.g. multi-agent system or decision support system.

3. Further work could be also made in examining some other criteria than misclassification criterion and considering situation with $\lambda \in [0, 1]$ (uncertain belonging to the relation).

4. In this work only static relational systems were considered. Special attention should be paid on dynamic relational systems. There are some research available already ([5, 9, 10] and especially [8]) but some more coherent survey and more general approach is needed. Dynamical relational descriptions should find a lot of applications in mechanics, economy and other fields. However, it should be underlined that in defining dynamical relational systems some other definitions has to be given like, for example, stability condition of relational system.

5. Very important issue seems to be knowledge integration and management. Nowadays it is one of the most important problem in modern computer science (e.g. [38]). It is easy to imagine a situation when knowledge given by different experts (or sources like institutions or agents in multi-agent networks) should be
CHAPTER 8. CONCLUSIONS

integrated. There is a variety of different approaches to integrate knowledge (e.g. [38]) but no such work was made for relational descriptions. And it seems that integrating RKRs should give very interesting and useful results.

6. Applying relation identification methods in relational complex systems (complex of operations). However, first definition of relational complex systems should be made for relational complex system identification. Some attempt was made (e.g. [54]) but it can be treated as a start point for further considerations. Besides, relational complex systems may have strong connection with relational knowledge integration.

7. Formulating general approaches for relation identification methods using soft variables. Methods presented in this work were based on random variable only but probably some more general methods can be proposed. Moreover, definition of relational system should be examined with definition of fuzzy relation [16]. Maybe some new definition of soft relation could be formulated.

8. Combining relational description with other knowledge description (e.g. rough relations, fuzzy relations, Takagi-Sugeno models with relations). It is even hard to predict what kind of further problems could be formulated in such hybrid representations.

The most interesting topics, from theoretical and practical point of view, are about relational knowledge integration (number 5 in above list), dynamic relational systems (number 4), identification of relational complex of operations (number 6) and general approach to knowledge representation with soft variables (number 7). Nevertheless, other issues are also worth of consideration.
Bibliography


[29] LAROSE D.T., *Discovering Knowledge in Data. Introduction to Data Mining*, John Wiley & Sons, New Jersey, 2005


[54] ŚWIĄTEK J., Identyfikacja złożonych systemów relacyjnych ("Identification of the complex relational systems"), in: Inżynieria Wiedzy i Systemy Eksper-
towe, eds. Z. Bubnicki, A. Grzech, Oficyna Wydawnicza PWr, Wrocław, 1993, pp. 281-288 (in Polish)


