Poznań University of Technology

Faculty of Computing Science and Management

Institute of Computing Science

MASTER THESIS

Knowledge Extraction from the Database of the Polish Registry of Congenital Malformations

Izabela Brzezińska

Supervisor: prof. dr hab. inż. Roman Słowiński

Poznań 2004

1. Table of contents

2. Introduction

Computer systems are commonly used nowadays in vast number of application areas, including banking, telecommunication, management, healthcare, trade, marketing, control engineering, environment monitoring, research and science, among others. Moreover there is a trend to use them anytime and anywhere. As a result a huge amount of data of different types (text, graphics, voice, video) concerning in fact all human activity domains (business, education, health, culture, science) is gathered, stored and available. For example there are collected business transactions, health records, account histories, results of scientific experiments, weather information, workload characteristics of computer network, logs of computer systems and users, web logs, etc. These data may contain hidden from a user interesting and useful *knowledge* represented (defined) by some non-trivial patterns, relationships, anomalies, rules, regularities, trends, and constraints.

It is worth noting that very valuable knowledge can also be included in information systems in which data are collected just for presentation or simple processing of independent entities and aggregates. As examples one can mention telecommunication billing systems, banking account systems and hospital information systems. These systems are completely sufficient and commonly accepted when considering some particular goals of information storing and processing, like invoice preparation, account checking or updating, diagnosis of disease or therapy ordering. However, it is easy to note that through carrying out global analysis on a whole group of entities (clients, patients) some new knowledge (rules, patterns, relationships, etc.) may be revealed.

With the growth of amount and complexity of the data stored in contemporary, large databases and data warehouses, the problem of extracting knowledge from datasets becomes a real challenge, increasingly difficult and important. This problem is a central research and development issue of *knowledge discovery* that generally is a non-trivial process of looking for new, potentially useful and understandable patterns in data [15]. However, it is subjective whether the discovered knowledge is new, useful and meaningful since it depends on the application.

The knowledge discovery process is composed of many steps that lead from raw data collection to the new knowledge [61]:

- getting familiar with the discipline to be analyzed, identifying accessible knowledge and users aims,
- choosing data connected with the aims of the process,
- data preprocessing,
- choosing the task and algorithms for knowledge discovery,
- knowledge extraction from the dataset,
- interpreting and evaluating of the discovered knowledge,
- preparing the knowledge for further usage.

It is important to stress the role of a user in the process of knowledge discovery. He needs to have a firm understanding of the analyzed discipline, so that he can make a reasonable and correct decision of dataset for the analysis, decide on preferable computer representation form of knowledge (e.g., decision rules, decision trees, clusters, multi-dimension regression models, contingency tables [32]), choose algorithms for knowledge extraction according to particular dataset and aims of the analysis. Therefore, often a coalition of domain experts (e.g., medical experts) and computer scientists is made to assure deep discipline knowledge as well as firm understanding of methodological and practical computational aspects of knowledge discovery process itself.

According to [15], knowledge discovery is an iterative and interactive process. Once the discovered knowledge is presented, the users can enhance its evaluation measures, select new data etc. in order to obtain different, more accurate results. Often, even partial results require expert's evaluation, which might bring modifications to initial specifications and force next process iterations.

The knowledge discovery from data is done by induction. It is a process of creating patterns (hypothesis, generalizations) which are true in the world of the analyzed data. Those patterns can take different forms of knowledge representation like decision rules or decision trees. However, it is worth mentioning, as Karl Popper did, that *one cannot prove the correctness of generalizations of specific observations or analogies to known facts, but can refute them.*

Knowledge discovering can be seen in perspective of [61]:

- prediction or
- description.

Prediction is a form of data analysis that concerns predicting future or unknown values of attributes on the basis of available data. In particular, the aim of prediction can be to predict assignment of objects to certain classes (categories) on the basis of knowledge coming from analyzing objects that have been classified in the past.

Description is a form of data analysis that consists in automated discovery of previously unknown patterns describing the general properties of the existing data and presenting it to the user in a clear form enabling further interpretations.

The data which is the input for knowledge discovery process is a set of objects (also called cases, instances, samples or examples) described by a vector of values of attributes. Attributes taking quantitative values are called *numerical* and attributes with qualitative domains are referred to as *nominal*.

With respect to particular, predefined goals of knowledge discovery one can distinguish many general approaches, where the most important ones are : classification, clustering, association analysis, characterization and discrimination.

In the thesis, attention is focused on the first of the above-mentioned approaches. In classification the goal is to find a concise, formal classification mechanism (model), called *classifier*, which for each objects described by a vector of condition attributes assigns this case (maximally) correctly to an appropriate class stating a value for so called dependent (decision) attribute. To meet this goal, first, the classifier is built by analyzing a dataset of training objects described by condition attributes and one known dependent (decision) attribute. More precisely, each training object is vector of condition attribute values (i.e., a set of <attribute-value> pairs) with associated class (decision attribute). Then, the resulting model is verified estimating its prediction accuracy for a test set of objects with know decision attribute. The test objects are randomly selected and are independent of training objects. Thus, for the testing set the value of the decision attribute is known but for the time of classification it is hidden. A misclassification occurs when the classifier is presented with an object and classifies it incorrectly i.e., predicts for the object a class different than its real, pointed by decision

attribute class. Each misclassification is treated as an error. *Accuracy* is a measure of classifier's performance defined as a percentage of correctly classified objects. In general, distinction among different types of errors may occur and in that case errors may not be of equal value. If distinguishing among error types is important, then a *confusion matrix* can be used to lay out the different errors. The confusion matrix lists the correct classifications against the predicted classifications for each class. The number of correct predictions is placed along the diagonal of the matrix [69].

There are many methods of estimating classifier's accuracy on new objects. Among the most common is the *k-cross fold-validation.* In this technique, the objects are randomly divided into *k* mutually exclusive partitions of approximately equal size. There are *k* iterations of the method and in each of them *k*-1 partitions are used as a training set (i.e., a set from which knowledge in a particular representation is extracted) and the one left out as a training set (i.e., a set on which accuracy is calculated). The average accuracy over all iterations is taken as the classifier's accuracy [69].

Finally, if the estimated by the test set accuracy of the classifier is acceptable, then this model can be used to classify future objects for which the value of the dependent attribute is missing or unknown.

There is a number of tools and methods for classifier construction: rough sets, instance-based learning and k-nearest neighbors, decision tree induction, statistical methods, etc. Each of them has its own way of dealing with missing values and inconsistencies in data. Moreover, the form of representation of the knowledge extracted from the data differs. Therefore, a comparison and evaluation of those methods, according to their predictive accuracy, interpretability and understandability on a real life dataset is an interesting task.

3. Aims and scope of the thesis

The Polish Registry of Congenital Malformations PRCM is the biggest and unique database in Poland keeping information about children who have been diagnosed as having congenital malformations in period between their birth and 2nd year of life, and children with congenital malformations who were born dead or unable to survive. It was founded as a scientific project ordered by the Polish Ministry of Health in April 1997 and since July 2000, it has been operating as part of the Government Programme of Monitoring and Primary Prophylaxis of Congenital Malformations in Poland. Fulfilling its objectives, the Registry provides the Polish Ministry of Health with important information necessary for healthcare management [73]. In June 2001 the PRCM joined the EUROCAT network, which is a European network of population-based registries for epidemiological surveillance of congenital anomalies. The Polish Registry of Congenital Malformations covers 72% of Polish population and currently contains in its database information about 32,000 children with congenital malformations described by a few dozen attributes [73]. It is the only source of such big amount of information about congenital malformations in Poland and one of rare registries of this kind in Europe.

In particular, the Polish Registry of Congenital Malformations is a unique source for data about children with Down syndrome suffering from congenital heart defects. The coexistence of Down syndrome and congenital heart defects has been examined and proved by many researches and reports. The incidence in the Polish Registry of Congenital Malformations of congenital heart defects among children with Down syndrome reaches 33%, which is much bigger than among the population of children without Down syndrome. The question of what the causes of such big incidence of congenital heart defects in population with Down syndrome are, remains open. The Down syndrome itself might be one of the main reasons, but it is also not out of the question that other factors like maternal and paternal age, birth weight, place of residence, etc., might have influence.

The aim of this work is an attempt to verify the latter possibility that is to extract knowledge in form of relationships between attributes like maternal and paternal age, birth weight, place of residence etc., and existence of congenital heart defects among

children with Down syndrome using data gathered in the database of the Polish Registry of Congenital Malformations.

This goal shall be attained by applying such different approaches to knowledge extraction as:

- rough set theory,
- instance based learning,
- decision trees induction,
- logistic regression.

In this context, an additional aim of the thesis is examination and comparison of different performance characteristics of the considered approaches.

The organization of the thesis is subordinated to the above aims and consists of 7 Sections. In particular, Section 4 introduces the medical basic facts about Down syndrome, congenital heart defects and their co-appearance. It gives few information about history and work of the Polish Registry of Congenital Malformations. Later on, it presents facts about previous analysis of congenital heart defect in Down syndrome and describes in details the dataset chosen for the knowledge extraction in this research.

Section 5 shows the application of rough set theory to knowledge extraction. First, it gives some methodological elements of the rough sets and rule based approach including classical rough set approach and generalization of rough approximation fro incomplete information systems. Later on, application of rough set theory to extraction of knowledge about congenital heart defects in Down syndrome is presented. This part talks about different experiments extracting knowledge in form of decision rules from the analyzed dataset and presents conclusions drawn from them.

Section 6 describes the application of instance based learning to knowledge extraction. The nearest neighbor methods and IBL1-3 algorithms are presented. Apart from that, the Section contains description of the course and conclusions of experiments applying instance based learning to knowledge extraction about congenital heart defects in Down syndrome.

Section 7 presents another approach to knowledge extraction - decision tree induction. It describes methodological elements of induction of decision trees, stressing the details of C4.5 implementation of tree induction algorithm. It also describes the

application of decision tree induction to knowledge extraction from the analyzed dataset, presenting carried out experiments and conclusions drawn from them.

Section 8 is dedicated to application of a particular statistical method called logistic regression to knowledge extraction. A methodological background is firstly presented, followed by description and conclusions of experiments extracting knowledge about congenital heart defects in Down syndrome.

Section 9 contains comparison of the rough set, instance based learning, decision trees and logistical regression approaches with respect to different performance characteristics.

Finally, Section 10 summarizes the thesis with a discussion on the completed work and possible future research.

4. Congenital heart defects in Down syndrome

4.1. Congenital malformations

A congenital malformation is a physical defect present in a baby at birth, irrespective of whether the defect is caused by a genetic factor or by prenatal events that are not genetic. In a malformation, the development of a structure of an organ (e.g. heart, brain, lungs) is arrested, delayed, or misdirected early in embryonic life and the effect is permanent [68].

Congenital malformations are a serious medical and social problem. They carry a high burden to affected individuals, their families and the community in terms of quality of life, participation in community and need for services. They are a significant cause of difficulties in procreation as congenital diseases of embryo often lead to its death. Moreover, they contribute strongly to mortality rate of newborns and infants. In case of congenital malformations among live borns, one third of them will suffer from mental and/or physical disabilities for their entire life [18].

Therefore, it is extremely important both for medical and social reasons, to carry out researches which aim to find dependencies and relationships between genetic as well as environmental factors (e.g., mother's age, maternity history, smoking during pregnancy) and congenital malformations. Discovering those dependencies and regularities, is a way of undertaking actions finding the causes of congenital malformations and perhaps limiting their consequences.

4.2. Down syndrome and congenital heart defects

Chromosomes are thread structures composed of DNA and other proteins. They carry the genetic information needed for the development of all the cell of the body. Human cells normally have 46 chromosomes arranged in 23 pairs. The medical test checking blood samples in order to determine the number and type of chromosomes is called a karyotype.

Down syndrome (also known as trisomy 21) is a genetic abnormality which is characterized by an "extra" 21st chromosome. All individuals with Down syndrome have extra chromosome 21 material. There are 3 genetic mechanisms for trisomy 21:

- non-disjunction
- translocation
- mosaic

Those mechanisms are responsible for the three types of Down syndrome respectively [5]:

- trisomy 21 (standard)
- translocation
- mosaicism

Human cells undergo two division processes called 'mitosis' and 'meiosis'. Mitosis is a process of cell division which results in the production of two daughter cells from a single parent cell. The daughter cells are identical to one another and to the original parent cell [64]. Meiosis is the type of cell division by which germ cells (eggs and sperm) are produced. This process involves a reduction in the amount of genetic material [64].

During those divisions some errors may occur and cause trisomy 21. One of them is called 'non-disjunction' and occurs when in meiosis division process one pair of chromosomes does not divide. In result, one cell will have 24 chromosomes and the other 22, where in normal situation each of them should have 23 chromosomes [35]. About 95% of all Down syndrome cases are caused by the event that the fertilized egg has three 21st chromosomes instead of two.

Another 3-4% of Down syndrome cases are due to another error in cell division process called 'translocation'. The point of translocation is that two divisions occur in separate chromosomes and usually the 14th and 21st chromosomes are involved. Due to such rearrangement some of the 14th chromosome is replaced by an extra 21st chromosome. In such case, even though the number of chromosomes remains normal, there is an extra part or whole 21st chromosome [35].

A small percentage of cases of Down syndrome is caused by error called ‘mosaicism’. In those cases people have some cell lines with normal set of chromosomes and some with trisomy 21.

The Down syndrome is verified by karyotype. An example of a male karyotype with trisomy 21 is shown in Figure 1.

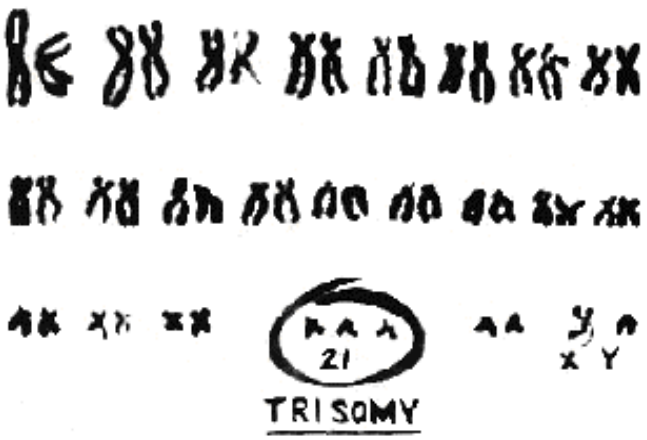

Figure 1. Example of a male karyotype with trisomy 21

Apart from distinguishing the three types of Down syndrome research has also shown the relation of Down syndrome with congenital heart diseases. The reported incidence is 40-50% [53].

Congenital heart diseases (CHD) are heart problems present at birth. There are many different types of congenital heart defects and more than one malformation may be present at the same time. The types of congenital heart defects include [63]:

- *Atrial septal defect (ASD)*
- *Ventricular septal defect (VSD)*
- *Atrioventricular septal defect (AVSD)*
- *Patent ductus arteriosus (PDA)*
- *Aortic Stenosis*
- *Pulmonary Stenosis*
- *Ebstein's anomaly*
- *Coarctation of the aorta*
- *Tetralogy of Fallot (TF)*
- *Transposition of the great arteries*

- *Persistent truncus arteriosus*
- *Tricuspid atresia*
- *Pulmonary atresia*
- *Total anomalus pulmonary venous connection*
- *Hypoplastic left heart syndrome*

4.3. Polish Registry of Congenital Malformations

In order to undertake any actions to limit the consequences of congenital malformations, let alone find its causes, it is necessary to thoroughly gather precise information about congenital abnormalities. This became one of the main reasons for the setting up of many registries throughout Europe collecting medical data.

4.3.1. *The history of the Polish Registry of Congenital Malformations*

The Polish Registry of Congenital Malformations PRCM was founded as a scientific project ordered by the Polish Ministry of Health and financed by the State Committee for Scientific Research in April, 1997. Since July 1st 2000, it has been operating as part of the Government Programme of Monitoring and Primary Prophylaxis of Congenital Malformations in Poland. Fulfilling its objectives, the Registry provides the Polish Ministry of Health with important information necessary for healthcare management [73]. In June 2001 the PRCM joined the EUROCAT network, which is a European network of population-based registries for epidemiological surveillance of congenital anomalies, started in 1979 and keeping in a standardized central database more than 160 000 cases of congenital anomaly among live births and stillbirths. More than 900 000 births per year in Europe are surveyed by 36 registries in 17 countries joined in EUROCAT.

4.3.2. *Area and population of Polish Registry of Congenital Malformations*

The Polish Registry of Congenital Malformations covers 73.6% of the area of Poland i.e. 230 091 km^2 with a population of 27 815 600 (72% of Polish population). According to data for the year 2001, about 265 000 births (live and still) a year are added to the registry's database, which makes about 72% of all births in Poland. Currently, the database contains information about 32,000 children with congenital malformations, born between January 1st 1997 and December 31st 2002 [73].

4.3.3. *Process of data acquisition*

The Polish Registry of Congenital Malformations has been built on the experience of many local registers kept in Poland through last decades. The PRCM registers children who have been diagnosed as having congenital malformations in period between their birth and 2nd year of life. Apart from that, children with congenital malformations who were born dead or unable to survive are registered. Each registration is made on the grounds of a one sheet printed form developed particularly for the purpose of gathering information for registration. The PRCM gathers all the registration forms sent by doctors from all over Poland and puts them into the PRCM's computer database [74].

4.3.4. *Database of the Polish Registry of Congenital Malformations*

The database keeping the information gathered on children with congenital malformations is run on an *Oracle 8.0 Database Server*. All the operations connected with the edition of database entries is done through an application made in *Oracle Forms Designer 6.0* [74]. The application has seven bookmark forms, each focused on different types of information gathered (personal data, description of pregnancy, previous pregnancies, malformations, death, parents, family). Several attributes have finite domains and therefore value lists are connected with them in the application.

Reports from the database are constructed by an application written in *Visual Basic for Applications* and can also be viewed and edited in *MS Excel* [74].

4.4. Analysis of congenital heart defects in Down syndrome

4.4.1. *Incidence of congenital heart defects in Down syndrome*

Mental retardation and muscular hypotonia are present in all children with Down syndrome, while the congenital heart defects (CHD) are observed in certain percent of children (16 to 62), depending on the population.

Table 1. Percentage of children with CHD in different world regions

Country/Region	Percentage of children with CHD among children with Down syndrome
Poland (PRCM)	33%
Dallas	52%
Atlanta	44%
Mexico City	58%
Rio de Janeiro	51%

The distribution of the types of congenital heart defects in children with Down syndrome is also different depending on population. According to the Polish Registry of Congenital Malformations [74], the most frequent heart defect was atrioventricular septal defect (AVSD) (33.8%), which is lower than the percentage among the US population of children with Down syndrome from around Dallas (45%). According to the Mexican authors [51], AVSD incidence in children with Down syndrome barely amounts to 14%. In the Rio de Janeiro population of children with Down syndrome the most frequent heart defect was ventricular septal defect (VSD) (51%). In children with Down syndrome in the PRCM, the VSD incidence was 17%; whereas in 41% of the

children a combination of VSD and atrial septal defect (ASD) was observed. A relatively large percent of children with Down syndrome in the PRCM had an isolated ASD (24.5%). The incidence of Fallot's tetralogy was much lower than in the Brazilian population (2.8% vs. 20%). The percentage of Mexican children with ASD, VSD and patent ductus atreriosus (PDA) was 90%, whereas in the children from the PRCM only 55.5%.

The incidence of types of congenital heart defects in children with Down syndrome in the PRCM is shown in Figure 2.

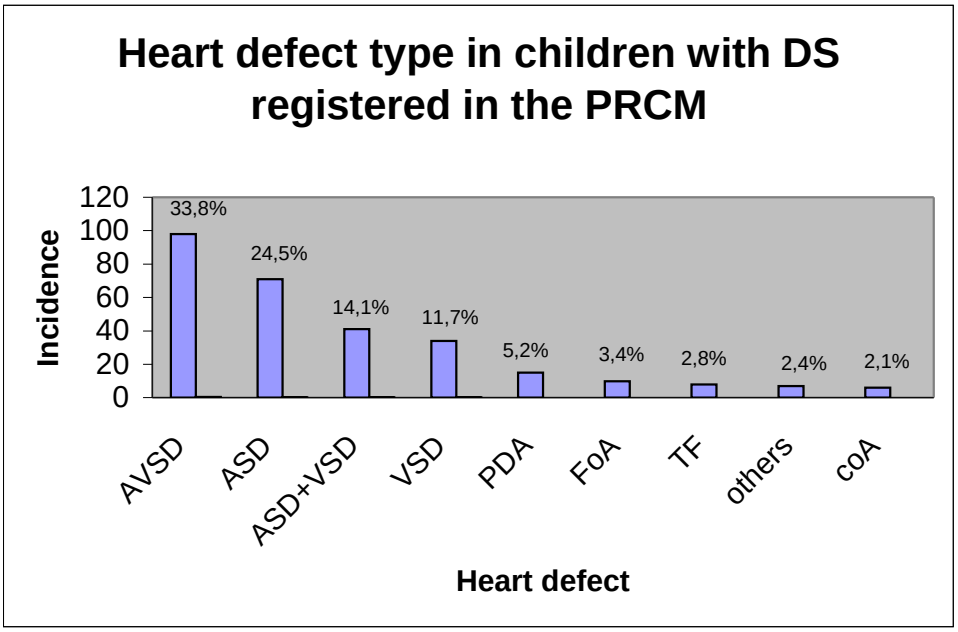

Figure 2. The incidence of types of CHD in children with Down syndrome registered in the PRCM.

4.4.2. *Factors influencing the incidence of congenital heart defects in children with Down syndrome*

4.4.2.1. *Genetic factors*

The recent studies have indicated that the congenital heart defect incidence in children with Down syndrome is conditioned by genetic factors (probably, the gene on

chromosome 21) and environmental factors. The extra chromosome material in Down syndrome can lead to instability in the process of morphogenesis, which makes the fetus vulnerable to teratogens.

The study of rare cases of patients with congenital heart defect and partial duplication of chromosome 21 made it possible to determine a congenital heart defect chromosomal region in Down syndrome. In patients with duplication of the region that is distal to 21q22 no heart defects were observed. In patients with duplication of the 21q22 region, heart defects were present in 50% of all cases.

Barlow [4] proposed a candidate gene localized on the chromosome 21, coding an adhesion molecule, expressed during the heart morphogenesis. He called it Down Syndrome Cell Adhesion Molecule. According to him, three copies of the gene were responsible for the presence of congenital heart defect in fetus with trisomy of chromosome 21.

Venugopalan [67] observed a much higher incidence of congenital heart defect in Down syndrome (60%) than in other populations. His study was carried out, however, on a relatively small group of children with Down syndrome (n=54).

4.4.2.2. *Environmental factors*

In the Alexandria study [42], the influence of environmental factors in early pregnancy on the incidence of congenital heart defects in infants with Down syndrome (n=514) was examined. The analysis concerned genetic, biological and environmental factors. Consanguinity of child's parents, consanguinity of mother's parents, taking antibiotics and hormonal contraceptives during pregnancy, and diabetes were observed to have influence on the incidence of congenital heart defects in the children with Down syndrome.

On the other hand, the studies by Loffredo [37] failed to confirm the influence of mother's diabetes on the incidence of complete AVSD in children with Down syndrome, despite the fact that mother's diabetic condition and taking antitussives increased the risk of isolated complete AVSD incidence.

The case control study in Dallas [65] focused on the influence of mother's illnesses, taken medication and psychoactive agents, and exposure to chemicals, between at least 3 months before the last period and 3 months after the last period. No significant influence related to the mother's age, income, parents' education, maternity

history, smoking, contraceptives, alcohol consumption during pregnancy, exposure to chemicals and intoxicants, on congenital heart defect incidence in children with Down syndrome was observed. The groups of women with diabetes, thyropathy, epilepsy, arthritis, hypertension, fever in the periconceptional period were rather statistically insignificant (5-10 women in each group). Only 15 mothers were taking antibiotics, and the study of this particular parameter seems to be rather difficult as well. Overall the entire group under study was not large (171 children with Down syndrome in total).

The study in Atlanta [28] focused on 243 cases of Down syndrome. Mother`s race or mother's age had no influence on congenital heart defect or AVSD incidence in children with Down syndrome. It has been, so far, the only population-based study.

On the basis of the results of the conducted studies the following conclusions can be made:

1. Studies in different countries use different methodology and diagnostic standards of congenital heart defect detection and classification of children into the control group.
2. The number of studied cases in total is fairly insignificant.
3. Due to the differences between the results of the studies by various authors, a comprehensive population-based study of Down syndrome children with and without congenital heart defect is required.

4.4.3. *Dataset description*

4.4.3.1. *Data preprocessing*

Preprocessing is an important phase of the knowledge extraction process in which dataset in being transformed so that it meets requirements of further analysis. Before applying methods of pattern recognition to data, it is crucial to put the dataset into proper input form. Actions dealing with: extraction of data from different sources, "cleaning" the noisy data, solving the problem of incompleteness and inconsistencies, discretization etc. can be taken into account as parts of preprocessing.

Extraction of data from different sources and eliminating duplicates

The dataset used in this analysis all comes from one database, therefore none of the possible problems appearing when taking data from different sources had occurred. In some cases, however, duplicates occurred i.e., there was more than one entry of the same child in the database. Such situation took place if some additional information about child's malformation was gathered after child's first registration to Polish Registry of Congenital Malformations. Following entries formed kind of a history of the patient. For the purpose of the analysis, such histories had to be put into one entry as no duplicates were allowed. The process of putting some entries into one i.e., eliminating the duplicates, was done using *MS Excel* after data extraction from the database.

Dealing with noisy data

Informational noise is understood as mistakes in attribute values. They can be accidental (a person entering the data to database had made a mistake) or can be caused by the alternations in the measured attribute. Fortunately, no alternations have been made to analyzed attributes. Thus, the noise could be caused by humans either in the phase of filling the printed registration form or later, at the time of entering the data from the registration forms into the database. Thanks to equipping the application for entering the data with *value lists* on all attributes which have finite domains, the latter could be avoided. Therefore, noise could be brought into the data only at the registration phase and are unfortunately impossible to identify after they have been entered to the database. One can only count on vigilance of the experts entering the data to the Polish Registry of Congenital Malformations.

Inconsistencies in data

Inconsistent data is understood as any two objects which are described by the same values of all condition attributes but are in different decision classes. The problems of inconsistent data were not tackled in the preprocessing phase but left to be dealt with by particular analysis methods.

Incompleteness of data

Any object that does not have values of all attributes is considered as incomplete. In the database of the Polish Registry of Congenital Malformations there are all together 893 cases of children with Down syndrome. The value of the decision

attribute is given for each of those cases, however, some examples are incomplete due to missing values of condition attributes. It has been observed that 867 cases have not more than one value of condition attribute missing. The dataset has been reduced by leaving out cases with more than one attribute value missing. The reduction of the dataset by 2,9% has brought improvement of the quality of classification (see chapter 3) and therefore all further analysis have been performed on the slightly reduced dataset. These missing values were not tackled in the preprocessing phase but left to be dealt with by particular analysis methods.

Handling numerical values

Attributes with real number or integer domains are called numerical attributes. Such large domains are difficult to handle by human perception. Rules or other knowledge representations are easier to understand and remember, and therefore more useful, if they are based on as small domains as possible. Therefore, it is required to apply *discretization techniques*. Discretization is a process of changing the numerical attributes into discrete, ordinal ones. It is done by dividing the original domain of the numerical attribute into a certain number of cut-points and assigning certain symbolic codes to those cut-points. In general, no discretization method is optimal for all situations. There are 4 numerical attributes in the analyzed dataset: *birth weight*, *fetal age, maternal age*, *paternal age.* For the first two numerical attributes discretization compatible with the medical standards have been used. In case of two other numerical attributes (i.e., *maternal age, paternal age*) a version of a minimal entropy method with a stopping condition referring to a maximum number of intervals per discretized attribute has been applied. The specific cut-points have been given in descriptions of particular attributes below.

4.4.3.2. *Description of attributes used in the analysis.*

All the data comes from the Polish Registry of Congenital Malformations and describes 867 cases of children with Down syndrome, among which 290 (about 34%) have a congenital heart disease (CHD). An exemplary subset of the dataset is shown in Table 2. Each case is described by ten condition attributes and one decision attribut telling whether the child has a congenital heart disease (CHD=yes) or not (CHD=no). All attributes with their possible values are listed in Table 3 and described with details

later on in this chapter. The distribution of objects among the two decision classes (CHD=yes / CHD=no) is highly imballanced with the favour of CHD=no class. Some of the objects have missing values on condition attributes denoted by '?'-value.

Table 2. Exemplary subset of the dataset

object	place of residence	sex	cytogenetic examination	fetal age	birth weight	maternal age	paternal age	obstetrical history	smoking father	smoking mother	CHD
1	country_side	male	non_disjunction	0	1	0	1	no	no	no	no
2	very_large_city	female	non_disjunction	?	1	0	1	no	no	no	no
3	very_large_city	male	non_disjunction	0	1	0	0	no	no	no	yes
4	large_city	male	non_disjunction	0	1	1	1	yes	no	no	no
5	country_side	female	non_disjunction	1	1	1	1	no	no	no	no
6	country_side	female	non_disjunction	1	1	0	1	no	no	no	no
7	country_side	male	non_disjunction	1	1	1	1	yes	no	no	no
8	country_side	male	non_disjunction	?	1	1	1	no	no	no	no
9	country_side	male	translocation	1	1	0	0	no	no	no	no
10	small_city	male	non_disjunction	1	1	0	0	no	no	no	yes
11	country_side	male	non_disjunction	?	0	1	0	no	no	no	no
12	country_side	female	non_disjunction	1	?	0	0	no	no	no	no
13	small_city	female	non_disjunction	0	1	0	1	yes	no	no	yes
14	large_city	male	non_disjunction	0	1	0	0	yes	no	no	no
15	small_city	female	non_disjunction	0	0	0	0	no	no	no	no
16	?	female	non_disjunction	1	1	0	0	yes	no	no	no
17	small_city	female	non_disjunction	0	0	0	1	no	no	no	no
18	small_city	male	non_disjunction	1	1	0	0	no	no	no	no
19	country_side	male	non_disjunction	1	1	0	0	no	no	no	no
20	small_city	male	non_disjunction	0	1	0	0	no	no	no	no
21	large_city	?	mosaic	1	1	?	1	yes	no	no	no
22	country_side	female	non_disjunction	1	1	1	1	no	no	no	no

Table 3. Attribute names and their possible values

	attribute name	possible values				
1	birth weight	underweight	non-underweight	?		
2	fetal age	premature	non-premature	?		
3	sex	male	female	?		
4	maternal age	<38	>=38	?		
5	paternal age	<34	>=34	?		
6	obstetrical history	yes	no			
7	smoking mother	yes	no			
8	smoking father	yes	no			
9	place of residence	country side	small city	large city	very large city	?
10	results of cytogenetic examination (cytogenetic exam)	non-disjunction	translocation	mosaic		
11	CHD-decision attribute	yes	no			

The input dataset is defined by a mixture of numerical and qualitative attributes. Six of the condition attributes had originally nominal domains and four attributes had numerical (continuous) domains. Two of the numerical attributes have been discretized according to medical standards and two other according to results of a minimal entropy method with a stopping condition referring to a maximum number of intervals per discretized attribute. The condition attributes are as follows:

- birth weight
- fetal age
- sex
- maternal age
- paternal age
- obstetrical history
- smoking mother
- smoking father
- place of residence
- results of cytogenetic examination

Birth weight

Birth weight is a nominal attribute with two possible values: *underweight*, *non-underweight*. Originally in the database, the attribute has a numerical domain with unit of measure being one gram, and allowing all greater than 0 values. The original attribute has been discretized into a two-value domain, assigning the value *underweight* to cases with the birth weigh smaller than 2500 grams and value *non-underweight* to all others. This discretization corresponds well with the medical standards. Due to incompleteness on the birth weight attribute in the 24 cases, the domain of the attribute has been extended by adding a '?'-value to indicate the missing value on this attribute.

The distribution of values of this attribute according to the presence or absence of congenital heart defect is shown in Figure 3. The number of underweight children in the analyzed dataset reached 80%, showing significant imbalance in the distribution of values of birth weight. Only 19% of all cases are children with proper birth weight. There is 1% of values missing.

The distribution of cases with particular value between decision classes is similar to the distribution of cases in the whole dataset between decision classes. There are 34% of cases from the dataset that were assigned to CHD=yes class. From the subsets of children with underweight and without it, almost 33% and 37%, respectively, have congenital heart disease.

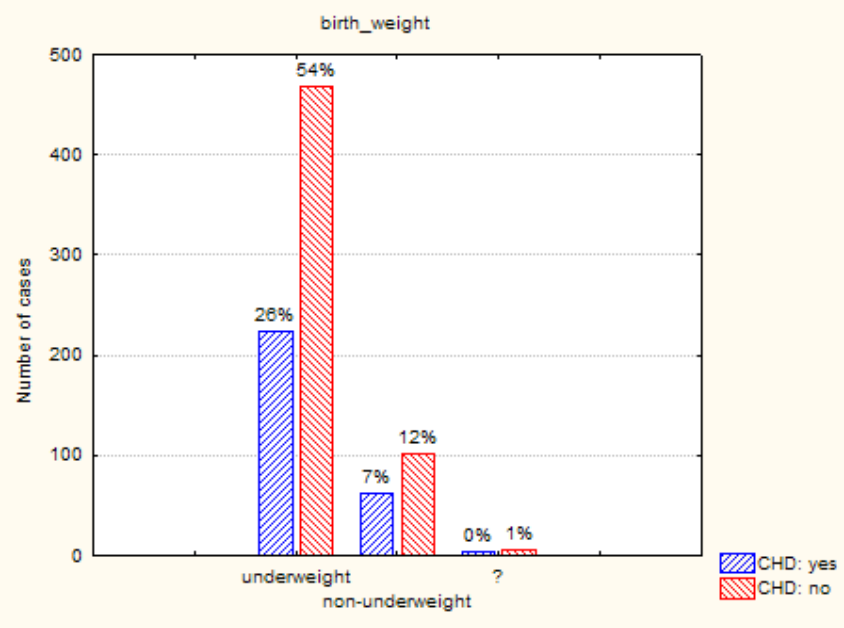

Figure 3. The distribution of values of *birth_weight* attribute categorized by decision attribute

Birth weight may be significant in a twofold way, and it is quite hard to interpret the role of this attribute. First, lower birth weight may be a result of a number of fetal factors, including the presence of CHD in the fetus, and it is not the lower birth weight which is a factor influencing fetal CHD incidence. At the same time, lower birth weight for maternal reasons may influence a lower survival rate among children with CHD.

Fetal age

Fetal age is an originally numerical attribute with values expressed in number of weeks of pregnancy, transformed into a nominal attribute. The possible values are: *premature baby* (corresponding to children born before the 38^{th} week of pregnancy) and *non-premature baby* (describing all children born in or after 38^{th} week of pregnancy). The bounds of discretization match medical standards. In 46 cases from the database, the fetal age was not given and therefore '?'-value had been added to the domain of this attribute.

The distribution of values of fetal age according to decision class is presented in Figure 4. 64% of case from the dataset were born prematurely and 33% were born in or after 38^{th} week of pregnancy. For 4% of children the fetal age was not obtained.

The distribution of cases with particular value between decision classes is similar to the distribution of cases in the whole dataset between decision classes as 33% of prematurely born children and 36% non-prematurely born ones have congenital heart disease.

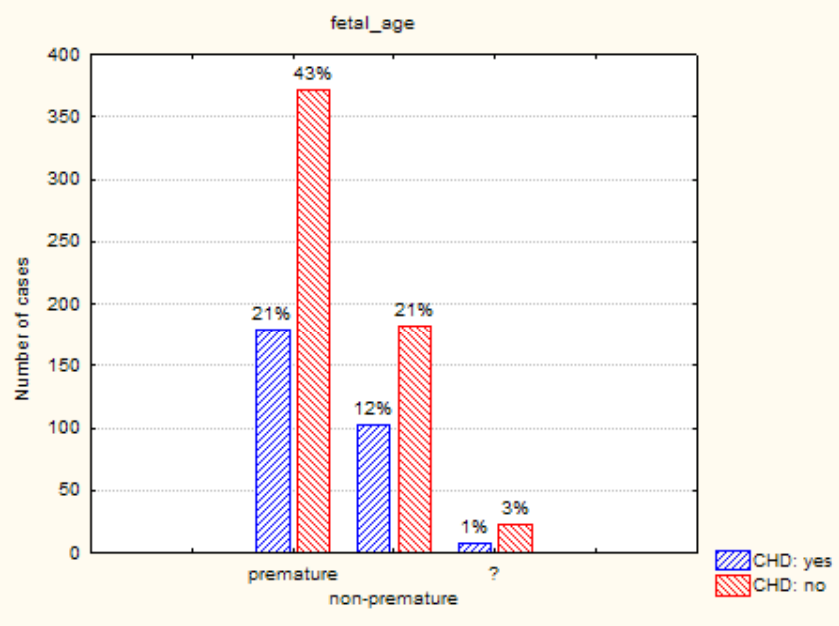

Figure 4. The distribution of values of *fetal age* attribute categorized by decision attribute

As in the case of birth weight, the role of this attribute is difficult to interpret. CHD in fetus may predispose to a preterm birth. There are also congenital heart defects such as FoA and PDA clearly related to prematurity.

Sex

Sex is a two value (*male*, *female*) nominal attribute. In 2 objects the sex of the baby was missing in database, so a third value '?' was added to the attributes domain.

The distribution of *male* and *female* values according to decision class is presented in Figure 5. The majority of children from the dataset are boys (56%). Since less than 1% of values is missing, girls constitute 44% of all cases. From congenital heart disease suffer 30% of boys from the dataset and 38% of girls. All together, 34% of children have heart malformations.

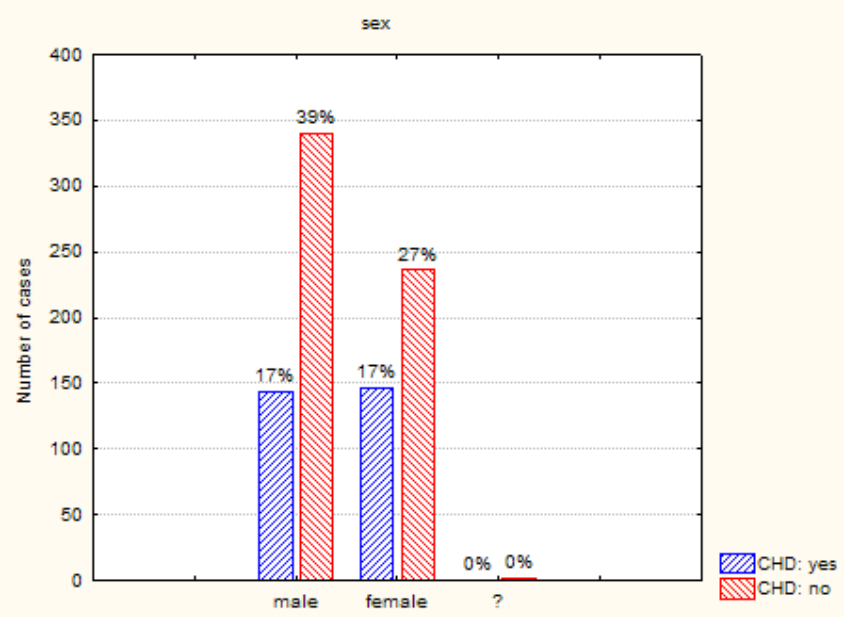

Figure 5. The distribution of values of *sex* attribute categorized by decision attribute

Overall, males are dominant in the children with Down syndrome. According to [34], mortality among Down syndrome children is higher in females, the major factor influencing the increase in mortality being CHD. It can be speculated that CHD could be a reason for fetus selection, and this would mostly concern females fetuses.

Maternal age

Maternal age is an originally numerical attribute with values expressed in number of years of mother at giving birth, transformed into a nominal attribute by discretizing it in order to minimize the entropy. Two ranges have been introduced: range 0 describing mothers under the age of 38, and range 1 for mothers 38 years old and older. The attribute domain also contains the '?'-value indicating missing values.

In Figure 6, it is shown that most (67%) of mothers of children from the dataset were younger than 38 years old. Almost 33% of them were at least 38 years old at the moment of birth giving. Less than 1% of gathered data was missing. Children with congenital heart disease were born to almost 36% of mothers from the first group and to 28% of mothers from the latter.

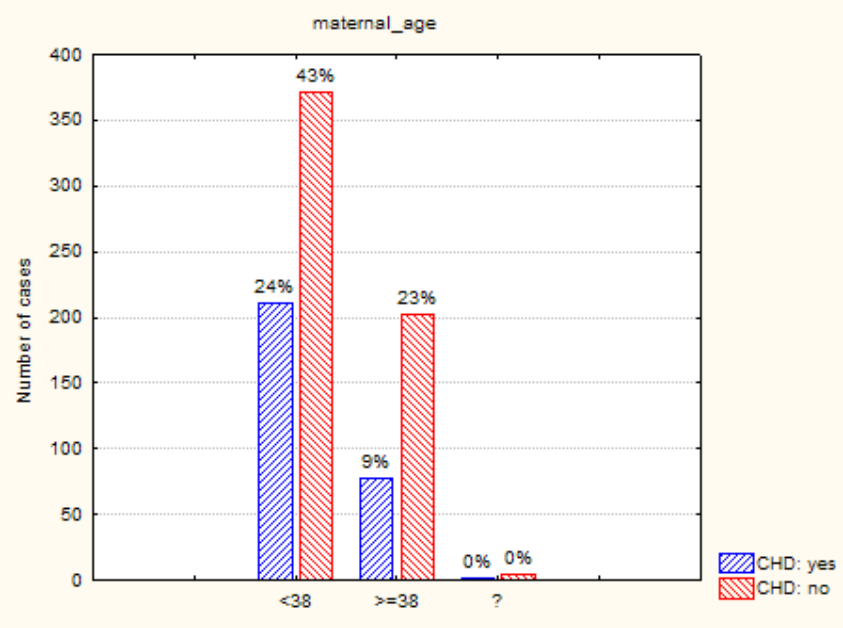

Figure 6. The distribution of values of *maternal age* attribute categorized by decision attribute

The analysis of this attribute is very complex. The advanced maternal age is an independent risk factor of Down syndrome incidence, thus a larger number of older mothers in the studied group can be observed. Maternal age is also linked to the fetal survival rate. Older women have a higher prevalence of conditions that could increase the risk of fetal death (diabetes mellitus, hypertension, *placenta praevia*, placental abruption). Cano [51] examined the survival rate of fetuses in women undergoing ovum donation from young donors (women over 40 years of age compared with women below 40 years of age). It turned out that older mothers had more difficulties in keeping pregnancies. The results suggest that the mechanisms responsible for normal functioning of fetoplacental unit are delayed in older patients. The proper delivery of substrate seems to be of crucial importance, and the failure of the uterine vasculature in older women may be responsible for this delay. An increase in miscarriage rates in older women might be related to the ability for vascularisation in older pregnant uteri, which brings about worse adaptability to the increasing hemodynamic demand of pregnancy. The older mother's age can be a risk factor in selection of fetuses with CHD. Therefore, more children with CHD will be born of young mothers, whose organism is able to keep pregnancy, despite the presence of CHD in the child.

Paternal age

Paternal age is an originally numerical attribute with values expressed in number of years of father at child's birth, transformed into a nominal attribute by discretizing it in order to minimize the entropy. Two ranges have been introduced: range 0 describing fathers under the age of 34, and range 1 for all older fathers. The attribute domain also contains the '?'-value indicating missing values.

The distribution of values of paternal age according to decision class is presented in Figure 7. The division line set at the age of 34, resulted in almost equal subsets: almost 50% of fathers were younger than 34 and 48% at the age of 34 or older. For almost 3% of cases, the age of the father was not obtained.

30% of fathers under 34, had children with congenital heart disease. This amount rose up to 38% in cases of older fathers.

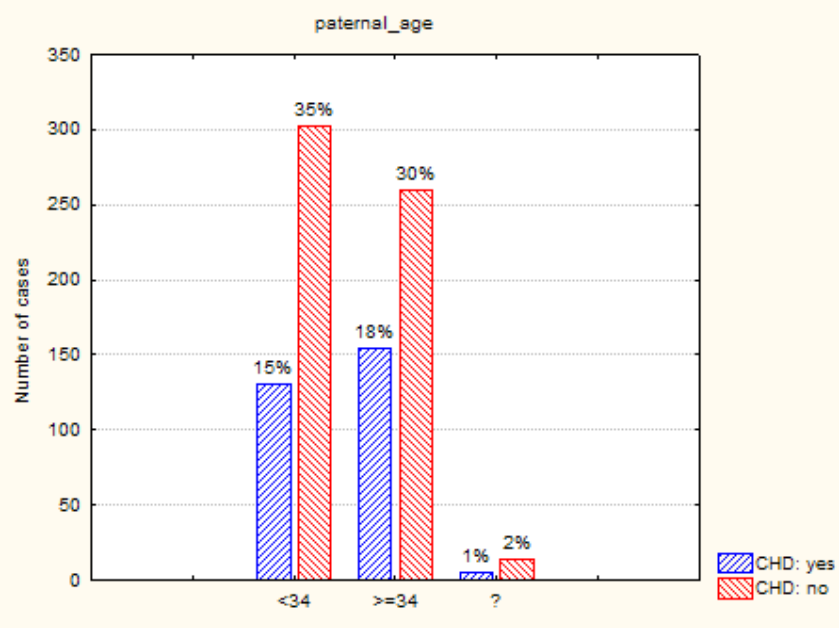

Figure 7. The distribution of values of *paternal age* attribute categorized by decision attribute

Paternal age is often inseparably linked to maternal age. In the majority of cases there is an insignificant age difference between both partners. Therefore, the seeming influence of paternal age on the incidence of CHD in children with Down syndrome can be, in fact, the effect of the influence of the mother's age on CHD incidence in the child. The future studies may concentrate only on the cases with larger age discrepancy between the partners.

Obstetrical history

Obstetrical history is an attribute aggregating two other attributes appearing in the PRCM's database: *miscarriages* and *fetal death*. The obstetrical history is an attribute with two possible nominal values: *yes* if the mother of the child has a history of miscarriages or fetal deaths, and *no* if no miscarriages or fetal deaths had happened. No missing values were observed therefore, there was no need to introduce '?'-value.

The distribution of values of this attribute according to the presence or absence of congenital heart defect is shown in Figure 8. The number of mothers without obstetrical history in the analyzed dataset reached 75%, showing significant imbalance in the distribution of values of this attribute. 25% of mothers had miscarriages of fetal deaths.

The distribution of cases with particular value between decision classes is as follows: 35% of cases without obstetrical history and 32% with it, were assigned to CHD=yes class.

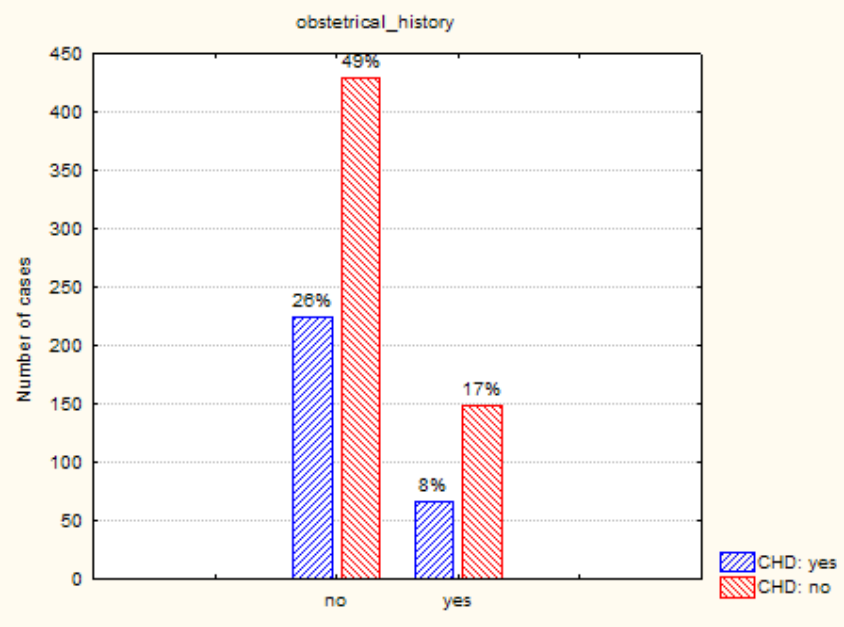

Figure 8. The distribution of values of *obstetrical history* attribute categorized by decision attribute

The influence of previous miscarriages and fetal deaths in mother on the incidence of CHD in fetus with Down syndrome is difficult to interpret as the PRCM lacks data on the underlying causes of miscarriages. Maternal predisposition to miscarriages (immunological and hormonal causes, congenital malformations of the

uterus, mother's illnesses) may be a factor limiting survival of the child with CHD. No precise analysis of the influence of this risk factor on CHD incidence in children with Down syndrome is possible without a detailed obstetrical history. As notifications to the PRCM are made by doctors from all over the country, we do not have full control over the precision of the obstetrical history related to them. Thus, the value of this parameter is rather doubtful.

Smoking mother, smoking father

These are nominal attributes with *yes* or *no* possible values. In Figure 9 and Figure 10 respectively, it is shown that over 98% of mothers and 96% of fathers did not smoke cigarettes during the pregnancy period. These two attributes are characterized by the greatest imbalance out of all attributes taken under consideration in his analysis. A bit over 33% of non-smoking parents had children with congenital heart disease.

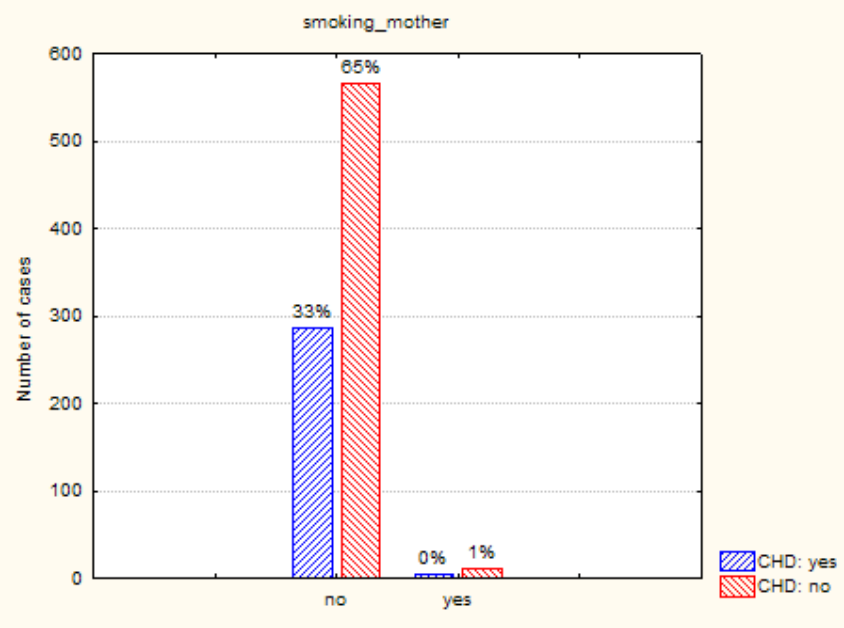

Figure 9. The distribution of values of *smoking mother* attribute categorized by decision attribute

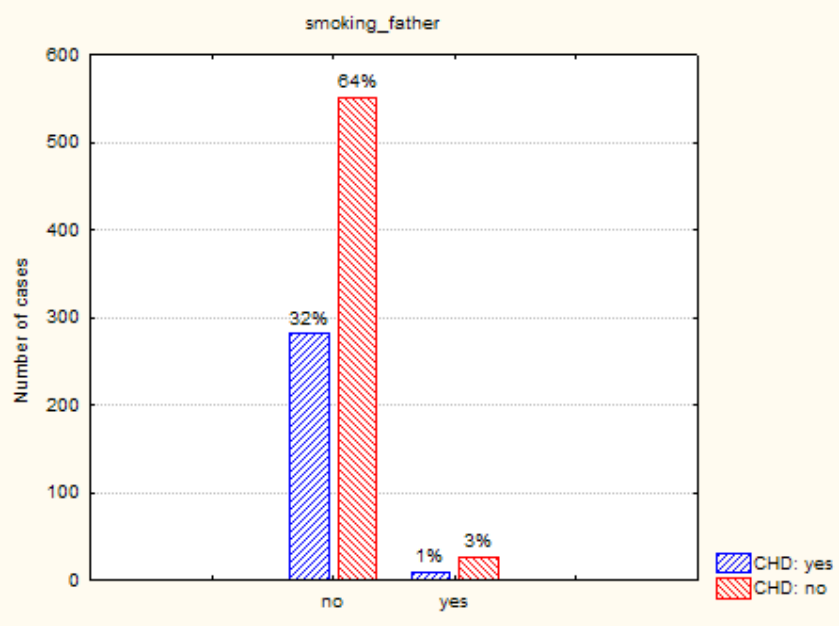

Figure 10. The distribution of values of *smoking father* attribute categorized by decision attribute

The epidemiological analysis of the studied group has shown that only 2% of mothers and 4% of fathers smoked cigarettes, which seems to be rather understated. The reason for it could be the initial lack of NO DATA AVAILABLE label in the PRCM database. The cases in which the notifying doctor had no access to data about father or mother smoking used to be marked in the database as NO SMOKING. The recent adding of the NO DATA AVAILABLE category to the database will allow a more precise articulation of this feature.

The effect of mother's smoking on the incidence of CHD in children with Down syndrome was examined by Fixler in [16]. He pointed out to insignificant differences in the CHD incidence between smoking and non-smoking mothers. However, according to Fixler, a serious limitation to this type of analysis (as well as to the analysis of exposure to other teratogens) is the fact that his study concentrated only on live births. What should be taken into account is the seemingly protective influence of smoking on a developing fetus with CHD in the course of Down syndrome: if teratogen dramatically increases mortality in fetuses with CHD, then the incidence of exposure among the live births will be the same in children with heart defect and without heart defect, or higher in children without heart defect than in those with heart defect.

Place of residence

Place of residence is a nominal attribute with values: *country-side, small-town, big-town, very-big-town* and a particular '?'-value indicating incompleteness of information in the database.

The distribution of values of this attribute according to the presence or absence of congenital heart defect is shown in Figure 11. In the dataset, there are 34% of cases coming from small city, 37% from country side, 9% from a very large city and almost 19% from large city. Almost 2% of values of place of residence are missing in the dataset. Among children from large cities, 42% suffer from congenital heart disease. Taking under consideration every other value of the attribute separately, it can be seen that, about 33% of cases of a particular value have congenital heart disease.

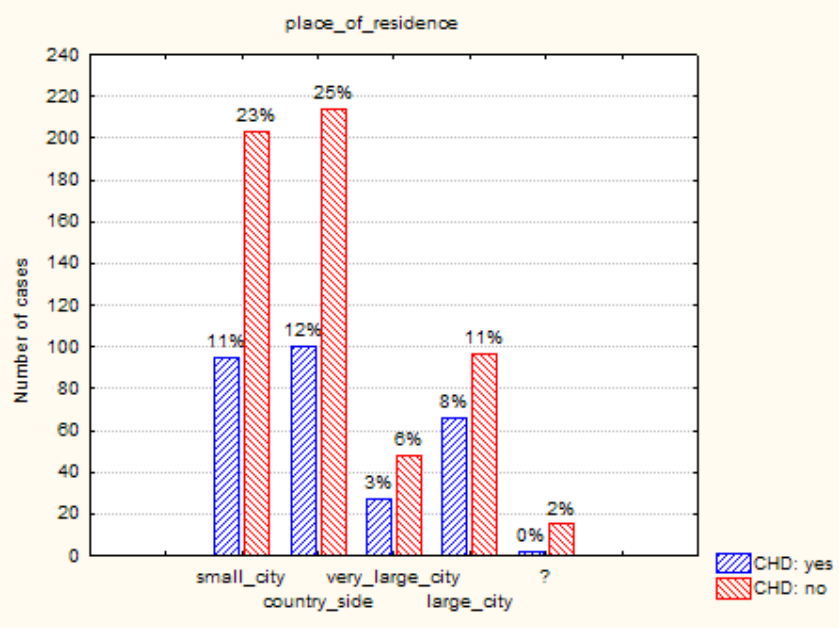

Figure 11. The distribution of values of *place of residence* attribute categorized by decision attribute

Place of residence may influence detection of CHD among children with Down syndrome. This is also linked to the older maternal age (older mothers in the studied group come more often from villages and small towns).

Results of cytogenetic examination

Results of cytogenetic examination is a nominal, completely filled-in in the database attribute with values: *non-disjunction, mosaic, translocation*. Figure 12 shows the distribution of values of results of cytogenetic examination according to the presence or absence of congenital heart defect.

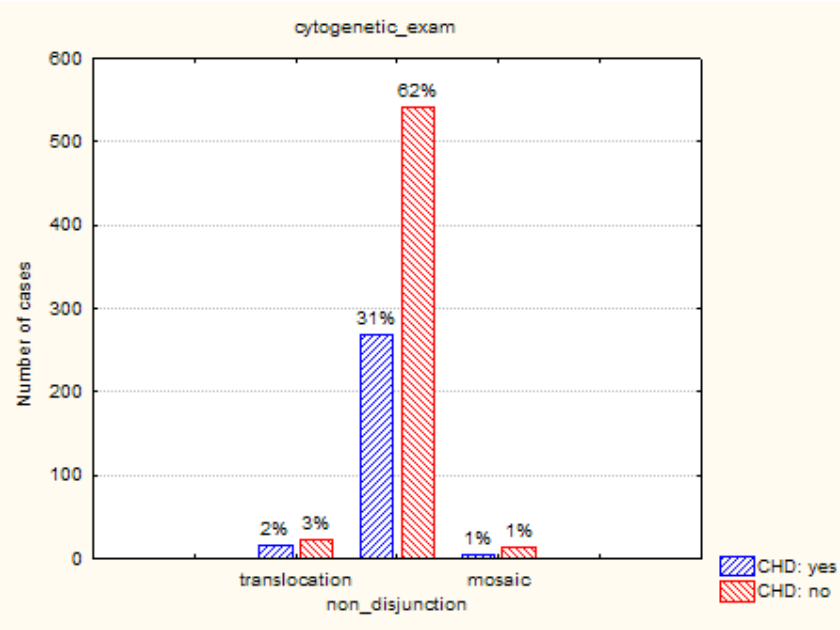

Figure 12. The distribution of values of *results of cytogenetic examination* attribute categorized by decision attribute

This is rather not an informative attribute as the most frequent (93%) chromosomal abnormality among children with Down syndrome is standard trisomy 21 caused by non-disjunction (in our group 810 children out of 867 had standard trisomy of chromosome 21). Children with translocation and mosaicism constitute only 5% and 2% of the dataset respectively. This is a significantly imbalanced in value distribution attribute. There were no missing values of this attribute in the dataset. In case of all values, the number of cases without congenital heart disease was larger than those with it.

4.4.3.2.1. *Missing values*

The dataset contains cases with missing attribute values. There are 4 condition attributes that had all values complete: *results of cytogenetic examination, obstertrical history, smoking mother* and *smoking father*. All other condition attributes had some

gaps. Table 4 and Figure 13 present the number of cases with missing values for each attribute. Clearly, the most incomplete attribute is *fetal age*, with 31 empty values, but it is only 3.58% of all cases in the dataset and therefore cannot disqualify this attribute from further analysis.

Table 4. Number of cases with missing attribute values

attribute	number of missing values	percentage of missing values
place of residence	17	1,96%
sex	1	0,12%
cytogenetic exam	0	0,00%
fetal age	31	3,58%
birth weight	10	1,15%
maternal age	5	0,58%
paternal age	19	2,19%
obstetrical hist	0	0,00%
smoking father	0	0,00%
smoking mother	0	0,00%

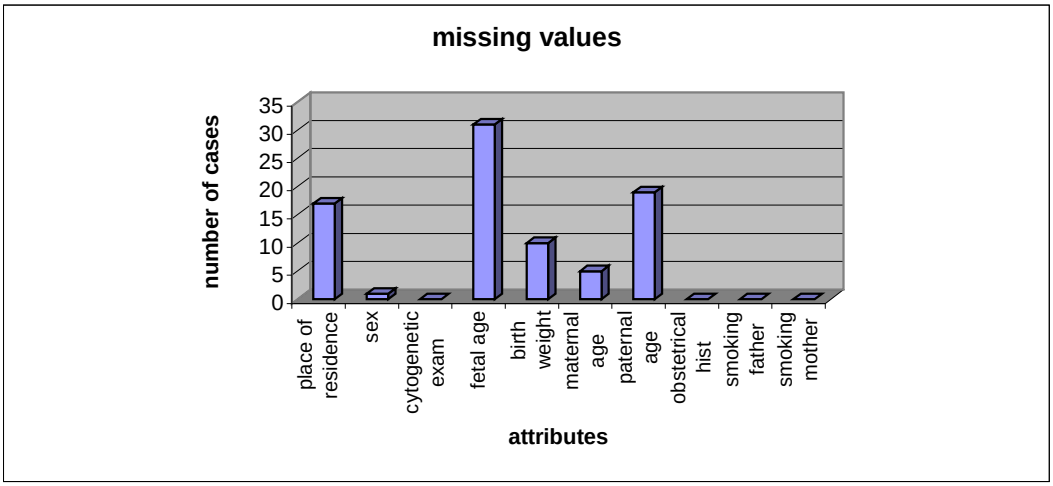

Figure 13. Number of cases with missing values on particular attributes

In Figure 14, distribution of objects with missing values among the whole dataset is shown. It is very interesting to see, how this distribution looks in each of the

decision classes. There are more incomplete objects in the class with congenital heart disease, but it is might be due to the fact that this class is larger in number of objects. Objects with missing values are rather equally distributed across the whole dataset.

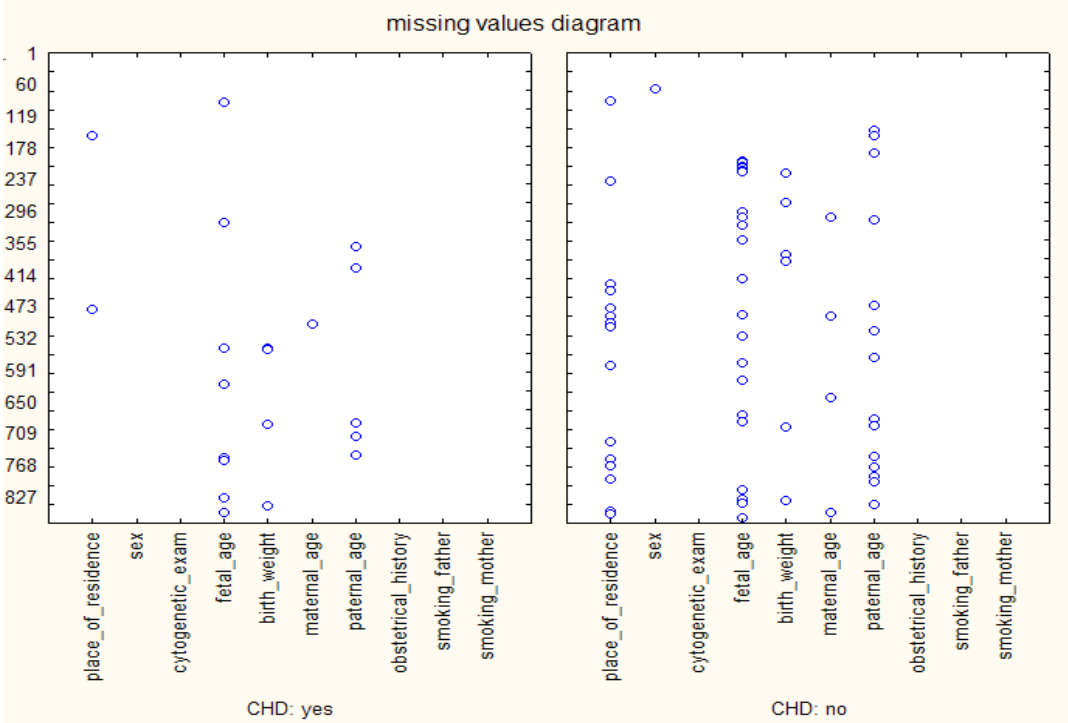

Figure 14. Distribution of cases with missing attribute values between decision classes of the dataset

4.4.3.2.2. *Crosstabulation and correlations*

Crosstabulation is a combination of two (or more) frequency tables arranged such that each cell in the resulting table represents a unique combination of specific values of crosstabulated variables. Thus, crosstabulation allows us to examine frequencies of observations that belong to specific categories on more than one variable. By examining these frequencies, we can identify relations between crosstabulated variables. Only nominal variables or variables with a relatively small number of different meaningful values should be crosstabulated.

In the analyzed dataset, eight out of ten condition attributes have a two-value domain. The *place of residence* and *results of cytogenetic examination* have five- and four-value domain respectively. Since discretization has been applied, all attributes are nominal and therefore ready as an input for the *crosstabulation*.

Table 5, Table 6 and Table 7 present more interesting results of crosstabulation. Cases with missing values are marked by '?'. Red-marked cells have counts >10.

Table 5. Crosstabulation results for *fetal age* and *birth weight*

birth_weight	fetal_age non_premature	fetal_age premature	fetal_age ?	Row Totals
non_underweight	494	171	27	692
underweight	51	110	4	165
?	6	4	0	10
All Grps	551	285	31	867

The results of crosstabulation for *fetal age* and *birth weight* show that there is a positive correlation between those two attributes since for children with underweight there is majority of children born prematurely over non-premature, and moreover, for non-underweight children there is majority of non-premature ones.

Table 6. Crosstabulation results for *maternal age* and *paternal age*

maternal_age	paternal_age 0	paternal_age 1	paternal_age ?	Row Totals
0	397	173	12	582
1	16	257	7	280
?	2	3	0	5
All Grps	415	433	19	867

The crosstabulation for *maternal age* and *paternal age* also shows a positive correlation between those attributes. That means that, partners tend to be of similar age.

Table 7. Crosstabulation results for obstetrical history and smoking mother

smoking_mother	obstetrical_hist no	obstetrical_hist yes	Row Totals
no	644	207	851
yes	9	7	16
All Grps	653	214	867

Obstetrical history and *smoking mother* are not correlated attributes. This may be due to strong imbalance in value distribution in *smoking mother* attribute.

Correlation is a measure of the relation between two or more variables. The measurement scales used should be at least interval scales, but other correlation

coefficients are available to handle other types of data. If the 2x2 table can be thought of as the result of two continuous variables that were (artificially) forced into two categories each, then the tetrachoric correlation coefficient will estimate the correlation between the two.

In Figure 15 there are diagrams presenting correlations between all possible combinations o pairs of attributes. Horizontal lines mean that there is no correlation between attributes according to the analyzed dataset. Diagonal lines appear in situations were a negative (falling line) or positive (rising line) correlation has been found. A correlation found through crosstabulation between *fetal age* and *birth weight*, and *maternal* and *paternal age* can be seen in the matrix as rising lines.

Figure 15. Correlation diagram matrix

5. Application of rough set theory to knowledge extraction

Rough set theory was introduced by Z. Pawlak [44] as a tool to deal with uncertainty and inconsistency in the analyzing information system. The granularity of the available data may be the cause of the inconsistency in objects' descriptions, but on the grounds of the rough set theory such inconsistencies can be dealt with. The theory is based on the assumption that having information represented in the form of attributes and their values on particular objects, it is possible to define an indiscernibility or other kind of relation between those objects [45], [47]. In the classical rough set theory the indiscernibility relation is used to build blocks of objects representing granules of knowledge about an universe. These granules are used in turn for approximation of sets or partitions of the universe that represent another knowledge about this universe.

5.1. Methodological elements of the rough sets and rule based approach

5.1.1. *Classical rough set approach*

5.1.1.1. *Information system*

An *information system* is a formal representation of the analyzed dataset and is defined as the 4-tuple

$$S = \langle U, Q, V, f \rangle ,$$

where U is a finite set of objects, Q is a finite set of attributes, $V = \bigcup_{q \in Q} V_q$ and V_q is a domain of an attribute q, $f : U \times Q \to V$ is a total information function, such that $\forall_{q \in Q, x \in U} f(x,q) \in V_q$.

The information system is in fact a finite data table which columns are labeled by attributes, rows by objects and the entry in column q and row x has the value $f(x,q)$. Each row in the table represents the information about an object in S, i.e. each object

$x \in U$ is described by a vector of attribute evaluations from Q. Such a vector is referred to as *description of x in S.*

In practice we are mostly interested in a special case of an information system called a *decision table,* where the set of attributes Q is divided into two disjoined subsets C and D ($C \cup D = Q$). The set C is a subset of *condition attributes,* and the set D contains *decision attributes* that determine the partition of U into decision classes.

5.1.1.2. *Indiscernibility relation*

The rough set theory is based on the observation that objects may be indiscernible (indistinguishable) due to limited available information. Objects with identical *descriptions* are called indiscernible. The *indiscernibility relation* on U can be associated with every non-empty subset of attributes $P \subseteq Q$ and is defined as follows:

$$I_P = \{(y, x) \in U \times U : f(x, q) = f(y, q), \forall_{q \in P}\}.$$

We say that objects x and y are *P-indiscernible* by a set of attributes P if $y\ I_P\ x$.

The relation I_P divides the set U into blocks of *P-indescernible* objects, called *P-elementary sets*. The *P*-elementary set containing objects
P-indiscernible with $x \in U$ is referred to as $I_P(x)$ and defined as:

$$I_P(x) = \{y \in U : y\ \mathrm{I}_P\ x\}.$$

5.1.1.3. *Approximation of sets*

Let X be a non-empty subset of objects from U and P be a non-empty subset of attributes from Q.

An object $x \in U$ *belongs certainly to X* if all objects from the *P*-elementary set $I_P(x)$ also belong to *X, i.e.* $I_P(x) \subseteq X$. A set of all objects certainly belonging to X constitutes the *P-lower approximation* of X , defined as:

$$\underline{I}_P(X) = \{x \in U : I_P(x) \subseteq X\}.$$

An object $x \in U$ *could belong* to X if at least one object from the *P*-elementary set $I_P(x)$ would belong to *X, i.e.* $I_P(x) \cap X \neq \varnothing$. A set of all objects that could belong to X constitutes the *P-upper approximation* of X , defined as:

$$\bar{I}_P(X) = \{x \in U : I_P(x) \cap X \neq \emptyset\}.$$

The *P-boundary* (*P-doubtful region of classification)* is defined as the difference between the *P*-upper approximation and *P*-lower approximation of X:

$$BN_P(X) = \bar{I}_P(X) - \underline{I}_P(X)$$

The set $BN_P(X)$ is the set of elements which cannot be certainly classified to *X* using the set of attributes *P*.

With every subset $X \subseteq U$, an *accuracy of approximation* of set *X* by *P* can be associated. The accuracy of approximation of *X* by *P* takes values from the range <0,1> and is defined as follows:

$$a_P(X) = \frac{|\underline{I}_P(X)|}{|\bar{I}_P(X)|}.$$

The *quality of approximation* of $X \subseteq U$ by attributes from *P* represents the relative frequency of objects from *X*, correctly classified using attributes from *P*. It is defined as a ratio:

$$g_P(X) = \frac{|\underline{I}_P(X)|}{|X|}.$$

Rough approximation of a subset $X \subseteq U$ can be extended to partitions of U, in particular to classification of objects into decision classes.

Supposing that *CLASS* is a partition of *U* into *n* decision classes, *CLASS*={ $CL_1,...,CL_n$ }, and *P* a non-empty subset of attributes from *C*.

Then, *P*-lower and *P*-upper approximation of *CLASS* is defined as:

$$\underline{I}_P(CLASS) = \{\underline{I}_P(CL_1),...,\underline{I}_P(CL_n)\},$$

$$\bar{I}_P(CLASS) = \{\bar{I}_P(CL_1),...,\bar{I}_P(CL_n)\}.$$

On such bases, the *quality of classification* can be defined as the ratio of all *P*-correctly classified objects to all objects in the decision table:

$$g_P(CLASS) = \frac{\sum_{i=1}^{n} |\underline{I}_P(CL_i)|}{|U|}.$$

A *reduct* is a minimal subset of attributes ensuring the same quality of classification as the entire set of attributes. More than one reduct can exist for one information system. The idea of reducts is interesting as it allows dataset size reduction. It is tempting to keep only those attributes that preserve approximation of classification and remove those that cannot worsen the classification.

A *core* is an intersection of all reducts in the information system.

5.1.2. *Generalization of rough approximations for incomplete information systems*

The classical rough set approach requires the information table to be complete. Since many practical problems are characterized by incomplete information, a generalization of the classical approach was needed to allow the information tables to have some empty cells. Below, we present an approach for dealing with missing values in rough approximation proposed by Greco, Matarazzo and Słowiński [20],[21], [22].

5.1.2.1. *Incomplete information system*

The *incomplete information system* is defined similarly to the information systems with the only change of adding to the set V a special symbol $*$ to indicate a missing value. This augmentation also introduces changes to the definition of the information function itself. Therefore the incomplete information system is defined as:

$$S^* = \langle U, Q, V^*, f^* \rangle$$

where $V^* = V \cup \{*\}$, $f^* : U \times Q \to V^*$, and $\forall_{q \in Q, x \in U} f^*(x,q) \in V_q \cup \{*\}$.

5.1.2.2. *Cumulative indiscernibility relation*

The *indiscernibility relation* from the classical approach was substituted by the cumulative relation on U associated with any non-empty subset P of Q:

$$I_P^* = \{(y,x) \in U \times U : f^*(y,q) = f^*(x,q) \vee f^*(y,q) = * \vee f^*(x,q) = *, \forall q \in P\}.$$

Thus, y and x are P-indiscernible if for any $x, y \in U, y I_P^* x$.

The *P*-elementary set containing objects *P*-indiscernible with $x \in U$ is defined as:

$$I_P^*(x) = \{y \in U : y\, I_P^* x\}.$$

5.1.2.3. *Cumulative approximations*

Let X be a non-empty subset of objects from U and P be a non-empty subset of attributes from Q. The following definitions can be introduced:

The *P-lower approximation* of *X*:

$$\underline{I}_P^*(X) = \{x \in U_P^* : I_P^*(x) \subseteq X\}.$$

The *P-upper approximation* of *X*:

$$\overline{I}_P^*(X) = \{x \in U_P^* : I_P^*(x) \cap X \neq \emptyset\}.$$

The *P-boundary* of X:

$$BN_P^*(X) = \overline{I}_P^*(X) - \underline{I}_P^*(X).$$

The *accuracy of approximation* of set *X* by *P*:

$$a_P^*(X) = \frac{|\underline{I}_P^*(X)|}{|\overline{I}_P^*(X)|}.$$

The *quality of approximation* of $X \subseteq U$ by attributes from *P*:

$$g_P^*(X) = \frac{|\underline{I}_P^*(X)|}{|X|}.$$

Rough approximation of a subset $X \subseteq U$ can be extended to partitions of U, in particular to classification of objects into decision classes.

Let *CLASS* be a partition of *U* into *n* decision classes, $CLASS = \{CL_1,...,CL_n\}$, and *P* a non-empty subset of attributes from *C*.

Then *P*-lower and *P*-upper approximation of *CLASS* is defined as:

$$\underline{I}_P^*(CLASS) = \{\underline{I}_P^*(CL_1),...,\underline{I}_P^*(CL_n)\},$$

$$\overline{I}_P^*(CLASS) = \{\overline{I}_P^*(CL_1),...,\overline{I}_P^*(CL_n)\}.$$

The *quality of classification*:

$$g_P^*(CLASS) = \frac{\sum_{i=1}^{n} |\underline{I}_P^*(CL_i)|}{|U|}.$$

The generalized approach takes the form of classical situation if the information table is free of empty cells. Due to the fact that the analyzed information system had missing values, the generalized approach had been used in all the work.

5.1.3. *Decision rules*

5.1.3.1. *Definition of a decision rule*

A decision table can be seen as a set of learning examples which enable induction of *decision rules* [61].

The decision rule is a logical statement:

if *COND,* ***then*** *DEC*

where *COND* is the condition part of the rule (i.e. the complex, which is a conjunction of elementary conditions called selectors on particular attributes) and *DEC* is a decision part (i.e. a disjunction of assignments to decision classes driven by the decision attributes) [70].

If the information table is consistent, rules are induced from the decision classes. Otherwise, they can be generated from approximations of decision classes. As a consequence to that, induced decision rules are categorized into *certain* and *approximate* ones depending on the used lower and upper approximations, respectively [21], [22].

Rules are usually formed in the way that they contain the operator '='. Then the rule can be expressed as:

if $f^*(x, q_1) = r_{q_1}$ ***and ...*** $f^*(x, q_p) = r_{q_p}$**,**

then *x* is assigned to $CL_{t_1} or ... or CL_{t_k}$.

where $q_i \in C$, $r_{q_i} \in V_{q_i}$ for i=1,2,...p and $CL_{t_j} \in CLASS$ for j=1,2,..k

If *k*=1, then the rule *r* is *exact*, otherwise it is *approximate*.

The rule *r covers (exactly)* the object $y \in U$ if *y* satisfies (exactly) the condition part of *r*. A set of objects from *U* that match (exactly) the rule *r* is called a *cover of r* and denoted as $COV^*(r)$ ($COV^x(r)$). It is clear that $COV^x(r) \subseteq COV^*(r)$ [61].

5.1.3.1.1. *Minimal rule*

An exact decision rule

if $f(x,q_1) = r_{q_1}$ ***and ...*** $f(x,q_p) = r_{q_p}$ **,**

then *x* is assigned to CL_t.

is *minimal* if and only if there is no other rule

if $f(x,h_1) = u_{h_1}$ ***and ...*** $f(x,h_m) = u_{h_m}$ **,**

then *x* is assigned to CL_t.

such that $\{h_1,h_2,...,h_m\} \subseteq \{q_1,q_2,...,q_p\}$ and $u_w = r_w$ for all $w \subseteq \{h_1,h_2,...,h_m\}$.

An approximate decision rule

if $f(x,q_1) = r_{q_1}$ ***and ...*** $f(x,q_p) = r_{q_p}$ **,**

then *x* is assigned to $CL_{t_1} or ... or CL_{t_k}$.

is *minimal* if and only if there is no other rule

if $f(x,h_1) = u_{h_1}$ ***and ...*** $f(x,h_m) = u_{h_m}$ **,**

then *x* is assigned to $CL_{t_1} or ... or CL_{t_k}$.

such that $\{h_1,h_2,...,h_m\} \subseteq \{q_1,q_2,...,q_p\}$ and $u_w = r_w$ for all $w \subseteq \{h_1,h_2,...,h_m\}$.

5.1.3.1.2. *Minimal set of rules*

A set of decision rules is *complete* if:

- each $x \in \underline{I}^*_c(CL_t)$ supports at least one exact rule pointing at the class CL_t, *for each* $CL_k \in CLASS$
- each $x \in BD^*_c(\{CL_{t_1},...,CL_{t_k}\})$ supports at least one approximate decision rule pointing at classes $CL_{t_1},...,CL_{t_k}$, for each $\{CL_{t_1},...,CL_{t_k}\} \subseteq CLASS$.

A set of rules is *minimal* if it is complete and if removing any of those rules would make the set non-complete.

5.1.3.2. *Rule evaluation*

In order to introduce quantitative measures of rule evaluation, let us use the following contingency table presented in Table 8.

Table 8. Contingency table for the rule ***if*** *Cond* ***then*** *Dec*

	Cond	*~Cond*	sum
Dec	$n_{DecCond}$	$n_{Dec\sim Cond}$	n_{Dec}
~Dec	$n_{\sim DecCond}$	$n_{\sim Dec\sim Cond}$	$n_{\sim Dec}$
sum	n_{Cond}	$n_{\sim Cond}$	n

In Table 8 $n_{DecCond}$ is the number of objects covering both *Dec* and *Cond*, $n_{Dec\sim Cond}$ is the number of objects covering *Cond* but not covering *Dec, etc,* n_{Dec} is the number of all objects covered by the rule, n is the number of all objects.

Rules can be evaluated by several measures e.g.

- *relative strength of a rule,*
- *confidence* (also called *discrimination level),*
- *length of a rule* (also called *simplicity*),
- *total number of rules.*

The *relative strength* of a rule is a ratio of the number of positive examples covered by this rule to the number of all positive examples in the class, and can formally be defined as:

$$\mathrm{Re}\,lStr(Dec \mid Cond) = \frac{\mathrm{n}_{\mathrm{DecCond}}}{\mathrm{n}_{\mathrm{Cond}}}.$$

The *confidence* of a rule is a ratio of the number of positive examples covered by this rule to the number of all examples covered by this rule, and can be formally defined as follows:

$$Confience(Cond \mid Dec) = \frac{\mathrm{n}_{\mathrm{DecCond}}}{\mathrm{n}_{\mathrm{Dec}}}.$$

The *length* of a rule is the number of selectors (elementary conditions in the condition part of the rule).

The *total number of rules* is the number of all induced rules under particular conditions.

5.1.3.3. *Generation of decision rules*

Methods of generating decision rules can be divided into three groups:

- *minimal set of rules* covering all objects from a decision table,
- *exhaustive set of rules* containing all possible minimal rules for a decision table,
- *satisfactory set of rules* containing all minimal rules that satisfy additional requirements (e.g. minimum acceptable strength of rules) [70].

For the purpose of the analysis, the method of generating the satisfactory set of rules was used. It is implemented as the Explore algorithm in the ROSE2 environment.

Table 9. The Explore algorithm

```
-------------------------------------------------------------------------------------------------
      procedure Explore (
            Input
                  dec : decision, for which rules are generated
                  pos : set of positive objects
                  neg : set of negative objects
                  max_len: maximum length of generated rules
                  min_sup: minimum power of support of generated rules
            Output
                  rs : set of generated rules
      )
      begin
   1     rs = {}
   2     ss = { s | COV*(s) ∩ pos ≠ {}∧ COV*(s)≠ pos ∪ neg }
   3     for each selector s in ss do begin
   4           if |COV*(s) ∩  pos| < min_sup then
   5                 ss = ss – { s }
   6           else if |COV*(s) ∩ neg| = {} then begin
   7                 ss = ss – { s }
   8                 r = ‘if s, then dec’
   9                 rs = rs ∪ { r }
   10          end
   11    end
```

```
12    cq = form a queue with all remaining selectors s1,…,sn from ss
13    while cq ≠ {} do begin
14         f = remove the first complex from cq
15         h = the highest index of selector involved in c
16         cs = {}
17         for i = h + 1  to n do begin
18              if COV*(f) ∩ si ≠ {}then
19                   append f ∧ si to cs
20         end
21         for each complex c in cs do begin
22              if |COV*(c) ∩ pos| < min_sup or len(c)> max_len
      then
23                   cs = cs – { c }
24              else if |COV*(c) ∩ neg| = {} then begin
25                   cs = cs – { c }
26                   r = ‘if c, then dec’
27                   if r is minimal then rs = rs ∪ { r }
28              end
29         end
30         cq = cq ∪ cs
31    end
      end { procedure }
```

The Explore algorithm first finds all acceptable selectors that are characterized by a non empty intersection of their exact cover and the set of positive objects. At the same time, those selectors should not cover all the objects in the information table.

Each selector goes through checking if a single condition rule, satisfying the specific requirements, can be formed on its basis (lines 3-11). Only selectors covering only the positive objects are transformed into rules (line 8). They are removed (line 7) just like selectors with insufficient cover (line 4). All other selectors are put into a queue (line 12).

From line 13 begins a loop creating at least two-condition-long rules. In each pass, the first complex is taken from the queue and expanded by adding such selectors that have a non-empty intersection of their exact covers and exact cover of the complex.

Complexes that meet the specific requirements (line 22) and cover only the positive objects, form rules (line 26). Newly created rules are added to the resulting set

if they are minimal. The remaining complexes that fulfill the specific requirements but satisfy also negative objects are added to the queue for further processing in the loop.

5.2. Application of rough set theory to extraction of knowledge about congenital heart defects in Down syndrome

The analysis of congenital heart diseases in Down syndrome has been performed according to the following steps:

step 1: application of various preprocessing techniques to the dataset,

step 2: attribute selection and determination of attribute importance based on rough set theory,

step 3: induction of decision rule set,

step 4: discussion of results.

All rough set calculations have been performed by the means of *ROSE2* software developed in the Laboratory of Intelligent Decision Support Systems of the Institute of Computing Science, at Poznań University of Technology [52].

5.2.1. *Application of preprocessing techniques to the data*

Before applying methods of knowledge extraction to data, preprocessing had been performed. Following actions were taken:

- duplicates were eliminated by putting all database entries of the same child into one entry,
- dataset was reduced by 2,9% by expelling cases with more than one attribute value missing,
- numerical attributes were discretized in the following manner:
  bith_weigth: (-inf; 2500) <2500; +inf),
  fetal_age: (-inf; 38) <38; +inf),
  maternal_age: (-inf; 38) <38; +inf),
  paternal_age: (-inf; 34) <34; +inf).

The former two attributes were discretized according to medical standards, and for the latter two a version of a minimal entropy method with a stopping condition referring to a maximum number (i.e., 2)of intervals per discretized attribute was applied.

5.2.2. *Attribute selection and determination of attribute importance*

Attribute selection and identification of "the best" subset of attributes is connected with rough set approaches of looking for reducts and core of attributes. Unfortunately, calculating a core in the analyzed dataset of congenital heart malformations with Down syndrome, did not result in a proper subset of attributes. In other words, the only reduct contained all ten condition attributes. Therefore no expected gain of dataset size reduction was obtained.

5.2.3. *Induction of decision rule set*

At this point of the analysis process it was aimed to extract from dataset some hidden information regularities represented in a form of interesting and useful to medical doctors decision rules. The obtained set of rules should facilitate understanding or interpreting collected medical experience.

The physicians were interested in extracting all rules that satisfy their requirements referring to the strength of decision rules. It has been performed by the means of the *Explore* algorithm. The parameter of the minimal relative strength was set to 7% both for the class with congenital heart disease (CHD=yes) and for the class without it (CHD=no) and the minimal discrimination of the rule was set to 75%, i.e. no rule with a ratio of number of cases matching it to number of cases covering only its condition part smaller than 75% has been accepted.. Having set the minimal relative strength, number of tests was made under changing the maximal length of the rule i.e. the number of elementary conditions in the condition part of the rule. The best quality of approximation has been obtained for the maximal rule length equal to 7. Under such conditions total number of 31 rules was induced, with the quality of approximation of 0,1880. They are presented in Table 10 sorted according to the strength of the rule.

The applied conditions on the rule strength were too severe to obtain rules from class CHD=yes. From medical point of view, correct assignment of children with congenital heart disease is more important as misclassifying unhealthy children to CHD=no class can have very serious results. Therefore, it was crucial to adjust the parameters of the rule induction algorithm in such a way that some rules from CHD=yes class would be induced. The total number of 4 rules was induced under the following conditions:

- the minimal relative strength = 2%
-the minimal discrimination = 75%
-the maximal length = 7.

These rules are presented in Table 11.

Table 10. Induced rules for the decision class of children without congenital heart disease

no.	rule	relative strength [%]	confidence [%]
1	(cytogenetic_exam=nondisjunction)& (birth_weight=1) & (maternal_age=1) **=>(CHD=no)**	28.60	75.00
2	(fetal_age = 1) & (birth_weight = 1) & (maternal_age = 1) **=>(CHD=no)**	21.49	75.61
3	(sex = M) & (maternal_age = 1) **=>(CHD=no)**	20.28	75.00
4	(residence = small town) & (cytogenetic_exam = nondisjunction) & (fetal_age = 1) & (birth_weight = 1) **=>(CHD=no)**	18.89	76.22
5	(residence = small town) & (sex = M) & (birth_weight = 1) **=>(CHD=no)**	18.89	75.17
6	(sex = M) & (fetal_age=1) & (birth_weight=1) & (paternal_age=1) **=>(CHD=no)**	18.37	75.18
7	(residence = country-side) & (cytogenetic_exam = nondisjunction) & (birth_weight = 1) & (paternal_age = 1) **=>(CHD=no)**	17.33	75.19
8	(residence = small town) & (fetal_age = 1) & (birth_weight = 1) & (obstetrical_history = no) **=>(CHD=no)**	16.64	75.59
9	(residence = country-side) & (birth_weight = 1) & (paternal_age = 1) & (smoking_father = no) **=>(CHD=no)**	16.64	75.00
10	(residence = small town) & (cytogenetic_exam = nondisjunction) & (fetal_age = 1) & (obstetrical_history = no) **=>(CHD=no)**	15.77	75.83
11	(fetal_age=1) & (birth_weight=1) & (obstetrical_history=yes)**=>(CHD=no)**	15.25	75.21
12	(fetal_age=1) &(maternal_age=1) & (obstetrical_history=no)**=>(CHD=no)**	14.73	75.22
13	(residence = small town) & (sex = M) & (fetal_age = 1) **=>(CHD=no)**	14.04	77.88
14	(residence = country-side) & (fetal_age = 1) & (birth_weight = 1) & (paternal_age = 1) **=>(CHD=no)**	13.34	75.49
15	(sex= M) & (birth_weight = 1) & (obstetrical_history = yes) **=>(CHD=no)**	13.00	75.76
16	(residence = small town) & (maternal_age = 1) **=>(CHD=no)**	12.48	75.00
17	(sex=M) & (paternal_age = 1) & (obstetrical_history = yes) **=>(CHD=no)**	11.44	75.00
18	(residence = small town) & (sex = M) & (paternal_age = 1) **=>(CHD=no)**	11.27	76.47
19	(residence = country-side) & (sex = M) & (paternal_age = 1) & (smoking_father = no) **=>(CHD=no)**	11.09	75.29
20	(residence = country-side) & (birth_weight = 1) & (maternal_age = 1) & (smoking_father = no) **=>(CHD=no)**	10.57	75.31

Table 10. Induced rules for the decision class of children without congenital heart disease- continuation

no.	rule	relative strength [%]	confidence [%]
21	(sex = M) & (cytogenetic_exam = nondisjunction) & (fetal_age = 0) & (paternal_age = 1) **=>(CHD=no)**	10.57	75.31
22	(residence = small town) & (cytogenetic_exam = nondisjunction) & (fetal_age = 1) & (paternal_age = 1) **=>(CHD=no)**	10.23	77.63
23	(residence = small town) & (fetal_age = 1) & (birth_weight = 1) & (paternal_age = 1) **=>(CHD=no)**	10.23	75.64
24	(sex = M) & (fetal_age = 1) & (obstetrical_history = yes) **=>(CHD=no)**	9.71	76.71
25	(residence = country-side) & (sex = M) & (birth_weight = 1) & (paternal_age = 1) **=>(CHD=no)**	9.01	76.47
26	(residence=country-side) & (maternal_age=1) & (obstetrical_history= no) **=>(CHD=no)**	9.01	75.36
27	(residence=country-side) & (birth_weight=1) & (obstetrical_history=yes) **=>(CHD=no)**	8.67	76.92
28	(residence = country-side) & (sex = M) & (paternal_age = 1) & (obstetrical_history = no) **=>(CHD=no)**	7.97	76.67
29	(residence = country-side) & (sex = M) & (fetal_age = 0) **=>(CHD=no)**	7.80	75.00
30	(sex = M) & (fetal_age = 0) & (birth_weight = 0) **=>(CHD=no)**	7.63	77.19
31	(sex = M) & (cytogenetic_exam = nondisjunction) & (maternal_age = 0) & (obstetrical_history = yes) & (smoking_father = no) **=>(CHD=no)**	7.28	75.00

Table 11. Induced rules for the decision class of children with congenital heart disease

no.	rule	relative strength [%]	confidence [%]
1	(residence=big_town) & (sex=F) & (birth_weight=0) & (paternal_age=0)& (obstetrical_history = no) **=>(CHD=yes)**	2.76	80.00
2	(residence = small_town) & (fetal age = 0) & (birth_weight = 1) & (maternal_age = 0) & (paternal_age = 1) & (obstetrical_history = no) & (smoking_father = no) **=>(CHD=yes)**	2.41	77.78
3	(residence = small_town) & (sex =F) & (fetal age=0) & (birth_weight=1) & (maternal_age = 0) & (paternal_age = 1) **=>(CHD=yes)**	2.07	75.00
4	(residence = big_town) & (sex = F) & (fetal age = 0) & (birth_weight=0) & (paternal_age = 0) **=>(CHD=yes)**	2.07	75.00

5.2.4. *Discussion of results*

Application of rough sets to knowledge extraction resulted in a set of 35 rules, which could facilitate understanding or interpreting of collected medical experience.

Although no unambiguous rule has been obtained in the analysis, it is clear that the induced rules can greatly enhance further epidemiological studies on congenital heart defects in children with Down syndrome. Without a doubt, they allow divagations concerning difficult links between the attributes, which might have seemed totally unrelated before. The rough set analysis is an excellent introduction to epidemiological studies on the role of individual parameters in congenital heart defects epidemiology in Down syndrome and the potential, interesting, relations between them.

The quality of approximation of the classification was low (0.1880). It means that the analyzed dataset is full of inconsistencies. Let us stress that this is a feature of the dataset, independent of the data analysis method used. Hence, unfortunately, the 10-fold cross-validation test of rules cannot be expected satisfactory as it is impossible to derive knowledge from ignorance. One of the reasons for such low quality of approximation can be the fact that there is no quality control over data at the phase of registration of a new case of congenital malformation. The paper registration forms filled in by physicians all over Poland sometimes contain information wrongly classified, interpreted or omitted. A kind of quality control should be introduced by substituting the paper registration forms by registration through a web side. This solution would ensure that no crucial information would be omitted. Applying value lists to as many attributes as possible would standardize physicians responses to registration form questions. Moreover, registration through a web site would eliminate the phase of inserting the information from the paper forms into the database, as the data could automatically be send to database. Many, difficult or even impossible to detect mistakes can be made at this phase if the data is inserted manually, not automatically. Switching to a computer solution at all data gathering phases would result in a better quality of the data and that could cause a better quality of approximation.

The gathered data is not an easy set to analyze. Unfortunately, the dataset did not allow to use all benefits of applying rough set theory - no proper reducts were found.

An experiment consisting in observing the quality of classification of different subsets of condition attributes was conducted. All possible subsets containing one or two condition attributes had 0.000% quality of classification. Moreover, there was one three-attribute subset that had the quality of classification just below 0.05%. The following attributes belong to that subset: *smoking mother*, *smoking father* and *birth weight*. This is the largest subset of attributes for which the quality o classification did not exceed the level of 0.05%. Therefore, we consider this subset as the subset of the least valuable and most noisy attributes.

There were 31 rules induced for the class without congenital heart disease (CHD=no). The best relative strength obtained was 28,6%. Three rules attained the level of 20% strength. The minimal strength boundary was 7%. The rules are mostly long (up to seven elementary conditions) but in general, their strength is satisfactory. Of course, none of the rules can be treated as a ready to use formula, but they can facilitate understanding or interpreting collected medical experience.

Only four rules were induced for the class of children with congenital heart disease (CHD=yes) and their relative strength was around 2%. In other words, the rule with the highest strength of matched only 2,76% of all cases from CHD=yes class i.e., the rule is supported by 8 objects. This result is rather poor, especially in the situation when distinguishing children from this class is more important from the medical point of view than children from the other decision class.

The dataset was characterized by a strong imbalance in the number of cases belonging to each of the decision classes (34% of cases belonging to class CHD=yes, 66% to class CHD=no). This could cause such an imbalance in the number and strength of induced rules for each decision class.

In order to evaluate the set of induced rules a *10-fold cross-validation* test was performed: the dataset was divided into 10 subsets and in each of 10 iterations one of the 10 subsets was used as the test set and the other 9 subsets were put together to

form a training set, then the average error across all 10 trials was computed. So far, we do not have access to objects that did not appear in the decision table during rule generation, therefore 10-fold cross-validation was applied.

A *confusion matrix* contains information about actual and predicted classification done by the classification system (i.e., set of rules). Confusion matrix was made to evaluate the performance of induced set of rules.

The results are shown in Table 12.

Table 12. 10-fold cross-validation results

Confusion Matrix (sum over 10 passes)		
	PREDICTED	
ACTUAL	CHD=no	CHD=yes
CHD=no	566	11
CHD=yes	287	3

Average Accuracy [%]		
	Correct	Incorrect
Total	65.64 (+-6.41)	34.36 (+-6.41)
CHD=no	98.18 (+-2.24)	1.82 (+-2.24)
CHD=yes	1.27 (+-1.08)	98.73 (+-1.08)

The total average accuracy exceeds 65%. It is very high (over 98%) for the decision class without congenital heart disease and, unfortunately, very low (over 1%) for the CHD=yes class. From medical point of view, it is much more important to receive high accuracy for the class with heart disease. The confusion matrix shows that the induced rules allowed to classify well only 3 out of 290 cases with congenital heart disease i.e. over 98% of examples from that class were incorrectly classified. This is an insufficient result, since a misclassification of ill patient costs much more that misclassification of patient without heart problems. Therefore, it is necessary to lead further research after working on dataset consistency.

5.2.4.1. *Clinical interpretation*

In most of rules (19/31) concerning children with Down syndrome without CHD the place of residence is usually a village or a small town. On one hand it might be the result of a lower exposure to the harmful effect of environmental pollution on pregnant women who live away from big urban centers. On the other side it could be connected with the limited access to echocardiographic diagnosis, thus with a lower detection rate of CHD in children whose mothers live away from the major academic centers.

In the rules concerning children without CHD, the male sex is dominant. This may be related to a higher survival rate of female fetuses with DS and coexisting CHD. Rule 3 seems to be very interesting; it indicates absence of CHD in male children of mothers in advanced age. This could be a confirmation of our hypothesis, with regard to the fact that older mothers carry to full term more seldom, in the cases in which the fetus has a CHD.

According to rule 30, boys are born more often without CHD, even if their birth weight is low and they are born prematurely. Do girls with low birth weight and born prematurely – due, for instance, to mother's disease – display a lower survival rate, if with CHD? We are not able to confirm that yet.

Rules 13 and 5 should be taken into closer consideration as well. They indicate that boys whose mothers live in small towns are born often without CHD, if they have proper birth weight and are born at term. We should wonder whether the absence of CHD predisposes to the proper birth weight and the birth at term or perhaps the fetuses - unexposed to mother's harmful factors – tolerate better the CHD incidence. The analysis into the rules in which absence of CHD correlates with previous miscarriages and fetal deaths in mother's obstetrical history will be very significant. Rule 11 indicates that infants with correct birth weight and born at full term, are born without CHD by mothers who have experienced miscarriages earlier. We assume these are mothers, whose risk of miscarriage is higher than in mothers who have never miscarried before. Is it then not a confirmation of our hypothesis that in the mothers with diseases constituting a potential risk of losing the fetus, the most prone to survive

are the strongest fetuses developing in a relatively correct manner (i.e. without CHD and with physical conditions allowing proper birth weight and birth at full term)? Moreover, this concerns boys, as illustrated by rules 24 and 15.

5.2.5. *Further experiments*

Selection is an operation that selects only some rows (cases) from the table (dataset) while discarding other rows.

Projection, on the other hand, is an operation that selects only certain columns (attributes) from the table (dataset) and discards the other columns.

The dataset obtained from the Polish Registry of Congenital Malformations underwent several selection and projection processes:

1. random selection of 290 cases from both decision classes
2. set of projections on 9 different condition attributes
3. set of projections on 8 different condition attributes

As a result we have obtained number of modified datasets to which application of rough set approach was applied. By the means of the Explore algorithm, sets of decision rules have been obtained from each modified dataset. The algorithm has been performed twice on each new dataset with the parameter of minimal strength of the decision rules firstly set to 7%, and then to 2%. We have observed the changes in the quality of approximation, number of attributes in the core and number of decision rules for each of the decision classes, depending on the change applied to the dataset.

5.2.5.1. *Experiment 1: selection*

This experiment has been carried out on a dataset in which the number of objects from both decision classes was equal. In the original dataset the decision classes were imbalanced, favoring the CHD=no class to which belonged 66% of all cases. This imbalance as well as low quality of some attributes caused extremely low quality of approximation in the original dataset. Moreover, only 4, very weak in strength, rules foe the class with congenital heart disease have been obtained.

Therefore, an experiment on a dataset with balanced decision class has been prepared. The balance was obtained by random selection of 290 cases from the CHD=no class and adding all (i.e., 290) cases from the CHD=yes class. The results of this experiment are presented in Table 13, for comparison the results obtained on the original dataset are shown in Table 14.

Table 13. Results after selection of dataset with a balanced in decision class

experiment number	quality of approximation	number of attrib in core	min rule strength [%]	number of rules	
				CHD=yes	CHD=no
1	0.2397	10	7	5	1
			2	75	35

Table 14. Results on the original dataset

quality of approximation	number of attrib in core	min rule strength [%]	number of rules	
			CHD=yes	CHD=no
0.1880	10	7	0	31
		2	4	80

Selection of the same number of cases from both decision classes resulted in over 27%-rise of the quality of approximation. Moreover, the number of rules obtained for the CHD=yes class exceeded the number for CHD=no class and from medical point of view, rules talking about presence of congenital heart disease are more valuable.

This experiment has shown that better results, in terms of quality of approximation and number of rules for congenital heart disease, can be obtained for a dataset with balanced decision class.

5.2.5.2. *Experiment 2: projection to 9 attributes*

Since no reducts had been found through the analysis of the original dataset, ten new datasets have been prepared, each containing only nine out ten original condition attributes, and rough set approach has been applied to them. Through projection to 9

attributes, 10 different datasets have been obtained and therefore ten rounds of experiment conducted.

The results of this experiment are presented in Table 15, for comparison the results obtained on the original dataset are shown again in Table 16.

Table 15. Results after projection - 9 attributes left

experiment number	number of condition attrib	omitted attrib	quality of approximation	number of attrib in core	min rule strength	number of rules	
						CHD=yes	CHD=no
1	9	smoking mother	**0.1765**	9	7	0	31
					2	**4**	75
2	9	smoking father	0.1603	9	7	0	27
					2	**3**	65
3	9	birth weight	0.1384	9	7	0	15
					2	0	43
4	9	paternal age	0.1373	9	7	0	19
					2	0	45
5	9	results of cytogenetic exam	0.1349	9	7	0	24
					2	**4**	60
6	9	maternal age	0.1338	9	7	0	23
					2	**2**	54
7	9	fetal age	0.1292	9	7	0	16
					2	**1**	42
8	9	sex	0.1223	9	7	0	16
					2	**1**	49
9	9	obstetrical history	0.1211	9	7	0	20
					2	**2**	41
10	9	place of residence	0.0704	9	7	0	12
					2	0	31

Table 16. Results on the original dataset

experiment number	number of condition attrib	omitted attrib	quality of approximation	number of attrib in core	min rule strength	number of rules	
						CHD=yes	CHD=no
1	10	-	0.1880	10	7	0	31
					2	**4**	80

Series of ten projections allowed to determine a kind of an attribute rank. Sorted in order of decreasing quality of classification Table 15, shows elimination of which attributes would have the least effect on quality of the classification. None of the examined ten subsets is as good, in terms of quality of classification, as the whole dataset (after all, no reducts were found) but it can be observed that eliminating *smoking mother* attribute worsens the quality of classification the least out of all attributes. That means that this attribute is the least informative, which could be caused by low quality control over the process of data acquisition and resulting from it high imbalance in value distribution (98% of all cases had value *no* on this attribute).

Low quality and imbalance in value distribution could also result in low informative abilities of attribute *smoking father*.

Attributes *birth weight* and *paternal age* were found to have correlation respectively with *fetal* age and *maternal age* and that may be one of the reasons for such high position of those attributes in the result Table 15.

5.2.5.3. *Experiment 3: projection to 8 attributes*

This experiment is continuation of the previous one. Four attributes which had the least decreasing effect on quality of approximation in Experiment 2 (*smoking mother, smoking father, birth weight, paternal* age) were eliminated in pairs from the original dataset. In the result five new datasets were prepared through the process of projection. By the means of the Explore algorithm, sets of decision rules have been obtained from each modified dataset and the results are presented in Table 17.

Table 17. Results after projection - 8 attributes left

experiment number	number of condition attrib	omitted attrib	quality of approximation	number of attrib in core	min rule strength	number of rules	
						CHD=yes	CHD=no
1	8	smoking mother, birth weight	**0.1269**	8	7	0	15
					2	0	42
2	8	smoking mother, paternal age	0.1257	8	7	0	19
					2	0	44
3	8	smoking father, birth weight	0.1130	8	7	0	13
					2	0	35
4	8	smoking father, paternal age	0.1119	8	7	0	17
					2	0	37
5	8	smoking father, smoking mother	0.0957	8	7	0	17
					2	2	33

Surprisingly, elimination of the two worst, according to Table 15, did not give the best result in terms of quality of approximation. Eliminating *smoking mother* and *smoking father* worsened the quality of approximation strongly, but only in that dataset rules for CHD=yes class were obtained. Since projection to 8 attributes resulted in extremely low quality of approximation and/or did not allow to obtain rules for the class with congenital heart disease, no further projections to even smaller number of attributes were performed.

6. Application of instance based learning to knowledge extraction

6.1. Nearest neighbor methods

The nearest neighbor method is one of the simplest learning methods. In order to give classification decision for a case, the method takes under consideration the case's nearest neighbor i.e., a case from a training set that is the closest to the case being classified in terms of a chosen distance metric of closeness. The distance metric expresses how similar the case being classified is to a case from the training set. The shorter the distance, the more similar the cases are. Among the most commonly used distance measures of closeness are:

- absolute distance
- Euclidean distance
- various normalized distances

Typically, the distance is calculated attribute (feature) by attribute and then summed up. When using the absolute distance, the absolute difference between features is summed up. In case of Euclidean distance, the difference between the values of each feature is squared and summed up for all features and the square root of this sum becomes the actual Euclidean distance.

The nearest neighbor classifier takes as an input the whole training set of samples, stores it and makes the classification decision for new samples using the training set. The classification decision is made in the following manner:

When a new case arrives, the nearest neighbor classifier calculates the distance between a new case and every case in the training set. If the training set contains a case exactly the same as the new case (i.e., a case the least possibly distant), then the classifier places the new sample into the training case's decision class. If, however, such a case is not found, then the nearest neighbor classifier pick the class of the closest training case [69].

The nearest neighbor method belongs to a family of methods known as the

k-nearest neighbors methods (*k-NN*). Instead of finding the single nearest neighbor, these classifiers look for k-nearest neighbors, where k is a constant. A new case is classified to the class that appears most frequently among the k-neighbors, and therefore, it is advisable to use an odd number of neighbors in order to eliminate possible ties.

The nearest neighbor method makes computationally no effort in learning from the samples, but the computation for predicting the classification of a new case is relatively large as the new case needs to be compared with every case in the training set. Therefore, the division for computation in learning phase and classification of new cases phase, is opposite to other classification methods like decision rules or decision trees. For those other methods, learning can be quite expensive, where as the classification and prediction on a new case typically involves a simple, computationally inexpensive matching step [69].

The weak point of the nearest neighbor algorithms is that they are non-incremental and their primary goal is to maintain perfect consistency with the initial training set. Although they summarize the data, they do not attempt to maximize classification accuracy on new cases and this ignores problems like noise, often met in the real life datasets. In order to overcome those problems, the idea of new group of algorithms called *instance based learning (IBL)* was presented. They are instead incremental and their goals include maximizing classification accuracy on subsequently showed instances [3].

6.2. Methodological elements of instance based learning

Instance based learning is, in general, a computer method that attempts to study solutions that were used to solve problems in the past in order to solve the current problem, by analogy or association. In this work, the focus was put on the group of instance based algorithms called IBL1-IBL3.

IBL1, IBL2, IBL3 were originally proposed and developed by David Aha and his collaborators [2] [3] [1]. In general, all of them are based on a variant of the *nearest neighbor algorithm*. The IBL algorithm takes as an input a set of training examples (traditionally called *instances*) and produces as an output a particular set of instances

called *concept description* (CD). An instance-based concept description includes a set of stored instances and, possibly, some information concerning their past performance during classification (e.g., their number of correct and incorrect classification predictions). This set of instances can change after presenting each training example [3]. Learning phase of such classifier finishes after having seen all instances from the training set and producing the final concept description. The phase of prediction of a classification for a new instance is done according to the nearest neighbor method with the concept description used as a set in which the nearest neighbor is searched for.
The learning phase of all IBL algorithms is composed of three parts:

- similarity function,
- classification function,
- concept description updater.

The *similarity function* computes the *similarity* between a training instance and the instances in the concept description.

The *classification function* receives as the input the computed similarities between the newly presented instance from the training set and each of the instances from the concept description and yields a classification for the new instance. For each instance from the training set, the yield classification is compared with the actual one, which is known since the new instance is a training instance, and on these grounds a classification accuracy (i.e., a ratio between the number of correctly classified instances and the number of all instances in the training set) of the concept description can be worked out. The concept descriptions' classification accuracy (*accuracy*) will be used to measure the performance of the instance-based learning algorithms.

The *concept description updater* maintains records on classification performance and decides which instance to add to the concept description and/or which to drop from the concept description.

In the prediction of classification for new instances from the training set or outside of it, IBL algorithms assume that *similar* (i.e. least distant according to some similarity function) *instances have similar classifications.* This leads to their local bias for classifying new instances according to their most similar neighbors' classification [1].

The similarity function used in IBL1, IBL2 and IBL3 is:

$$Similarity(C_1, C_2, P) = -\sqrt{\sum_{i \in P} Feature\_dissimilarity(C_1, C_2, i)}$$

where C_1 and C_2 are instances described by the set P of condition attributes,

and *Feature_dissimilarity(C_1, C_2, P)* is defined as:

an Euclidean distance $(C_{1i}\text{-}C_{2i})^2$ for numeric attributes

and for Boolean or nominal-valued attributes as:

$$Feature\_dissimilarity(C_1, C_2, i) = \begin{cases} 0 & if \quad C_{1i} = C_{2i} \\ 1 & if \quad C_{1i} \neq C_{2i} \end{cases}$$

It is assumed that missing values are maximally different from the presented value (i.e., they take the maximal possible for this type of attributes value of *Feature-dissimilarity*) and if they are both missing, their similarity is taken as 1.

6.2.1. *The IBL1 algorithm*

The IBL1 algorithm is the simplest instance-based learning algorithm. It is identical to the nearest neighbor algorithm with the exception that it normalizes its attributes' ranges, tolerates missing values, processes the examples incrementally and saves all processed training examples in the concept description (CD).

Table 18 presents the learning phase of the IBL1 algorithm. Later, new instances will be classified according to the nearest neighbor schema with concept description as the set from which the nearest neighbor is taken.

Table 18. The IBL1 algorithm

```
-----------------------------------------------------------------------
      procedure IBL1 (
            Input
                  TS : Training Set
            Output
                  CD : concept description
      )
      begin
```

```
1   CD = {}
2   if CD ≠ {}then
3      begin
4         for each x in TS do
5            for each y in CD do
6               Sim[y] = Similarity (x,y)
7         y_max = some y in CD with maximal Sim[y]
8         if class(x) == class(y_max) then
9            classification = correct
10        else classification= incorrect
11     end
12  CD = CD ∪ {x}
end { procedure }
```

The IBL1 algorithm first initiates an empty concept description (line 1). This makes the condition from line 2 **true** and therefore the first instance from the training set is added to the concept description (line12). This makes the condition from line 2 no longer **true,** and therefore each next instance from the training set goes through more complex processing (lines 3-11) before it is also finally added to the concept description (line12). For each not-first instance from the training set its similarity with every instance already belonging to concept description is calculated (line 6). Then, an instance from the concept description with the highest similarity value is chosen (line 7) and the classification proposed by this instance is compared with the real classification for the training instance (lines 8-10). Note, that these statements (lines 8-10) can be easily extended to count the number of correct and incorrect classifications. Such counters could be helpful when calculation the accuracy of classification.

The final output concept description contains all the instances from the training set.

As it was shown in several computational studies [2] [1]. this simple algorithm performs relatively well compared to other machine learning algorithms.

The weak side of this algorithm is its large storage requirements as it puts to the concept description all instances from the training set. That means that, the storage requirements will grow with the enlargement of the training set. It seems that the bigger the training set the better classification accuracy will be obtained, but those growing

storage requirements are a strong disadvantage. Therefore, a modification of IBL1 addressing that problem was proposed under a name of IBL2.

6.2.2. *The IBL2 algorithm*

The IBL2 algorithm was proposed to reduce the time needed to find the similar stored case matched and to reduce the storage requirements i.e. the number of instances remembered in concept description. The IBL2 learning phase, described in Table 19, is identical to IBL1's except that it saves only *misclassified instances*.

Table 19. The IBL2 algorithm

```
      procedure IBL2 (
            Input
                  TS : Training Set
            Output
                  CD : concept description
      )
      begin
   1   CD = {}
   2   for each x in TS do
   3      if CD == {}then
   4         CD = CD ∪ {x}
   5      else
   6         begin
   7            for each y in CD  do
   8               Sim[y] = Similarity (y,x)
   9            y_max = some y in CD with maximal Sim[y]
   10           if class(x) == class(y_max) then
   11              classification = correct
   12           else
   13              begin
   14                 classification = incorrect
   15                 CD = CD ∪ {x}
   16              end
   17         end
         end { procedure }
```

The algorithm at the beginning initiates an empty concept description (line 1). Like in IBL1, the first instance from the training set is added to the concept description unconditionally. Then each next instance from a training set goes through calculations (lines 6-17) in order to determine whether this instance should be added to the concept description or not. For each not-first instance from the training set its similarity with every instance already belonging to concept description is calculated (line 8), an instance from the concept description with the highest similarity value is chosen (line 9) and the classification proposed by this instance is compared with the real classification for the training instance (lines 10-16). Note, that these statements (lines 10-16) can be easily extended to count the number of correct and incorrect classifications. Such counters could be helpful when calculation the accuracy of classification. The not-first instance from the training set shall be join the concept description only if its real classification differed from the classification proposed by its most similar neighbor from the concept description.

The output is a concept description that contains those instances from the training set that have been incorrectly classified by its nearest neighbor.

The idea of saving only misclassified instanced has an intuitional justification by the fact that if the concept description is good enough to correctly classify a new instance from the training set, then there is no need to add such low-informative instance to the concept description. But in situations in which the classification of concept description failed, it is obvious that this instance from the concept description cannot classify the new training instance well. Therefore, the training instance is added to the concept description.

Empirical results obtained by D. Aha [2] [3] [1] show that IBL2 can significantly reduce IBL1's storage requirements only slightly decreasing classification accuracy.

IBL2's storage requirements can be potentially significantly reduced comparing to IBL1. The reduction is the greatest when none of the training instances is noisy. However, the storage requirements increase with the increase of noise level as noisy instances, with a high probability, will be misclassified and consequently saved. These noisy instances will probably misclassify the non-noisy examples and cause the growth of the concept description.

IBL2's classification accuracy drops more quickly than IBL1's when the level of noise in data increases (i.e., when the number of noisy instances increases). The reason

for that is that the noisy instances are almost always misclassified. Since IBL2 saves only a small percentage of non-noisy training instances, its saved noisy instances are more often used to generate classification decisions [3]. A solution aiming to tolerate noisy instances was presented as further modification of IBL2 called IBL3.

6.2.3. *The IBL3 algorithm*

The IBL3 algorithm is a noise-tolerant extension of IBL2 that keeps additional information about stored instances and tries to evaluate which of them will perform well during classification. IBL3 maintains a classification record i.e., the number of correct and incorrect classification attempts, with each saved instance. Moreover, IBL3 employs a significance test to determine which instances are good classifier and which ones are believed to be noisy (the former are added to CD, the latter discarded from it). The details of IBL3 learning phase are presented in Table 20.

Table 20. The IBL3 algorithm

```
-------------------------------------------------------------------------------------------------------
    procedure IBL3 (
          Input
                TS : Training Set
          Output
                CD : concept description
    )
    begin
 1   CD = {}
 2   for each x in TS do
 3      if CD == {}then
 4         CD = CD ∪ {x}
 5      else
 6         begin
 7            for each y in CD  do
 8               Sim[y] = Similarity (y,x)
 9            if ∃ {y in CD | acceptable(y)} then
 10              y_max = some acceptable y in CD with maximal Sim[y]
 11           else
 12              begin
 13                 i = a randomly selected value in [1|CD]
```

```
14                $y_{max}$ = some $y$ in $CD$ that is the $i^{th}$ most similar
                                  instance to $x$
15             end
16          if class($x$) == class($y_{max}$) then
17             classification = correct
18          else
19             begin
20                classification = incorrect
21                $CD$ = $CD$ ∪ {$x$}
22             end
23          for each $y$ in $CD$ do
24             if Sim[$y$] ≥ Sim[$y_{max}$] then
25                begin
26                   Update $y's$ classification record
27                   if $y's$ record is significantly poor then
28                      $CD$ = $CD$ – {$y$}
29                end
30       end
   end { procedure }
```

The algorithm at the beginning initiates an empty concept description (line 1). Like in IBL1 and IBL2, the first instance from the training set is added to an empty concept description unconditionally and calculation go on for next instances. Then similarity is calculated between the analyzed instance from the training set and all instances from the concept description (line 8). Next, the most similar neighbor is found (line 10) in the subset of *acceptable* neighbors (line 9) i.e., in the subset of those instances whose classification accuracy is significantly greater than its class's observed frequency (i.e., the percentage of processed training instances that are members of this class). If the classification proposed by the found neighbor is different from the real classification of the analyzed training instance (line 18), then the instance from the training set is added to the concept description (line 21). But once added to the concept description, the instance does not have to stay there. It shall be discarded from the concept description (line 28) if a significance test (line 27), which uses information gathered in the classification records, points it as believed to be noisy (i.e., its classification record is significantly poor).

For each training instance *x*, classification records are updated (line 26) for all saved instances that are at least as similar as *x*'s most similar ***acceptable*** neighbor (line 24). If none of the saved instances are acceptable (line 11), a random number i between 1 and the number of saved instances is taken (line 13) and the i most similar saved instances' classification records are updated (line 26). Based on the information in the classification record, instances accuracy is counted and instances are dropped from the concept description (line 28) if their accuracy is not significantly greater than theirs class's observed relative frequency.

Comparing a saved instance's accuracy with its class's observed frequency, decreases its sensitivity to skewed distributed concepts. Naturally, instances in concept description with high observed relative frequencies are expected to have relatively high classification accuracies as a relatively high percentage of its possible classification attempts will be for instances in its class. Analogically, instances from the concept description that have low observed relative frequencies are expected to have relatively low classification accuracies. IBL3 can more easily tolerate skewed concept distributions thanks to the mechanism of comparing instance's accuracy wit its class's frequency [3].

IBL3 assumes that the classification records of noisy instances will distinguish them from non-noisy instances as the former will have poor classification accuracies because their nearby neighbors in the instance space will invariably have other classifications [1]. This way, IBL 3 potentially will have less noisy instances in its concept description than IBL2 and therefore it is often referred to as a noise-tolerant extension of IBL2.

6.2.4. *Modifications of the IBL algorithms*

All IBL calculations were performed by the means of a slightly modified implementation of D. Aha's IBL algorithms developed at Poznan University of Technology by J. Stefanowski and S. Urbaniak. The introduced modifications extended original IBLs' abilities by adding more sophisticated approach to:

- handling nominal attributes
- working with data containing missing attribute values

6.2.4.1. *Modification to handling nominal attributes*

The original metric for nominal attributes was substituted by the ***value difference metric*** (VDM). It defines the distance between two nominal values x_i and x_j as:

$$dissimilarity(x_i, x_j, k) = \sum_{h=1}^{c} \left| \frac{C_{i,h,k}}{C_{i,k}} - \frac{C_{j,h,k}}{C_{j,k}} \right|$$

where k is the identifier of nominal attribute, c is the number of decision classes, $C_{i,h,k}$ is the number of learning instances that have x_i value of k attribute and belong to h-*th* decision class;

and $C_{i,k}$ is the number of all learning instances that have x_i value of k attribute.

This metric is always 0 if $i=j$. It is greater if the values of x_i and x_j discriminate well decision classes. The distance metric for numerical attributes is computed as before and both metrics are aggregated in the final similarity measure [62].

6.2.4.2. *Modification to working with data containing missing attribute values*

A technique based on substituting the missing value by the ***most common*** value of this attribute was proposed. Preprocessing of the input instances is needed to calculate the most common value. While calculating the most common value only instances belonging to the same decision class were taken into account [62].

The software turns those modifications on on the user's demand.

6.3. Application of IBL1-IBL3 to extraction of knowledge about congenital heart defects in Down syndrome

The analysis of congenital heart disease in Down syndrome has been performed using the IBL1-3 implementation extended by modifications handling nominal attributes and missing values. Missing values have been filled in by the most commonly appearing value from the particular attribute domain.

The parameters used to measure the performance of the classifiers were the *classification accuracy* (i.e., the concept description's classification accuracy, defined as the percentage of correct classification attempts) and the *storage requirements* (i.e., the size of the concept description, defined as the number of saved instances used for classification decisions). They were calculated during the 10-fold cross-validation process.

During the test we have manipulated with the following parameters: number most similar of neighbors when making classification decision (the *k* parameter for the k-NN algorithm) and turning on/off the modification for handling nominal values, in order to obtain the best classification accuracy on the analyzed dataset of children with Down syndrome and congenital heart malformations.

Number of tests has been carried out with changing parameters and the most informative results of IBL1 are presented in Table 21. For IBL2 and IBL3 algorithms the same tests have been performed but the result, in terms of classification accuracy obtained by IBL1 has not been beaten. In Table 22 are gathered the best results obtained by means of IBL2 and IBL3 algorithms.

Table 21. Results obtained from IBL1 algorithm

IBL1	**test 1**	**test 2**	**test 3**	**test 4**	**test 5**
no. of neighbors taken for classification	k=1	k=1	k=5	k=11	**k=11**
modification for handling nominal values	off	on	on	on	**off**
modification for handling missing values	on	on	on	on	**on**
no. of examples in concept description	867	867	867	867	**867**
classification accuracy	54.07 (+-1.46)	53.73 (+-2.27)	59.56 (+-2.15)	62.52 (+-1.06)	**64.50 (+-1.34)**

Table 22. Results obtained from IBL2 and IBL3 algorithms

type of algorithm	**IBL2**	**IBL3**
no. of neighbors taken for classification	k=11	k=11
modification for handling nominal values	off	off
modification for handling missing values	on	on
no. of examples in concept description	267	35
classification accuracy	56.27 (+-1.45)	54.96 (+-4.35)

In results obtained by IBL1 algorithm a growth of classification accuracy is observed with the increasing the number of nearest neighbors taken under consideration when making classification decision. Finally, classification accuracy has reached its highest value of 64.5% for *k*=11. Surprisingly, the results were better when the modification for handling nominal attributes was turned off. The modification did not bring the expected improvement in the result. The obtained classification accuracy, however, seems to be low. In a dataset where 64% of all cases belongs to one decision class, classification accuracy at 64.5% is not satisfying. If without any calculations, all cases would be classified as not having congenital heart disease, the classification accuracy would reach 64%. Therefore, obtained 64.5% is rather disappointing result. Moreover, classification accuracy in the decision class with congenital heart disease, presented in Table 23, is below expectations, as it is only 0.69%.

Table 23. 10-fold cross-validation results for IBL1 (test5)

Confusion Matrix (sum over 10 passes)		
	PREDICTED	
ACTUAL	CHD=no	CHD=yes
CHD=no	557	20
CHD=yes	288	2

Average Accuracy [%]		
	Correct	Incorrect
Total	64.50 (+-1.34)	35.50 (+-1.34)
CHD=no	96.53 (+-2.24)	3.47 (+-2.24)
CHD=yes	0.69 (+-0.58)	99.31 (+-0.58)

In tests 1-5 there were 867 cases taken as training instances and they all were added to the concept description according to IBL1's rule to take to concept description all the instances from the training set. Thus, no improvement on storage requirements was expected as long as working with IBL1.

For algorithms IBL2 and IBL3 analogical as for IBL1 tests have been carried out. Unfortunately, applying those algorithms did not result in improving classification accuracy. The best results obtained by means of IBL2 and IBL3 are presented in Table 22.

For IBL2 the highest classification accuracy was obtained in the same conditions (k=11, modification for special handling nominal values *off*) as the highest one for IBL1 and its value was 56.27%. Its a very low result, which shown that the strategy of remembering in the concept description only the misclassified instances, brought a decrease in classification accuracy by 8.23%. However, it resulted in reduction by over two thirds of the storage requirements.

In case of IBL3 algorithm, the classification accuracy dropped even more down to 54.96%. That was the best result obtained by IBL3 and it was done for k=11 and the modification for special handling of nominal attributes turned off. However, this algorithm, brought as expected the biggest reduction of storage requirements. In concept description there were only 35 instances. This shows that even though IBL3 did not

exceed the best classification accuracy of IBL1, its results comparing with IBL2's are very good. IBL2 needed 267 instances in concept description to reach 56.27% classification accuracy, while IBL3 kept only 35 instances in concept description and its classification accuracy dropped only by 1.31%.

All in all, obtained classification accuracies were disappointingly low and this brings us back to the discussion about dataset quality. Perhaps, introduction of quality control in the phase of data acquisition, would reduce the amount of noise and thereby also inconsistency in the dataset which could give an effect of higher classification accuracies. However, it also undeniable that the analyzed dataset is difficult with its imbalances of value distribution, inconsistencies and missing values. This also has an effect on achieved classification accuracies. One cannot forget, that the analyzed data is a real life dataset, in which inconsistencies might not only be caused by noise but also be the effect of existence of conflicting observations in the real life.

6.3.1. *Further experiments*

Since IBL1 algorithm had performed the best in terms of classification accuracy on the analyzed dataset, it was also additionally applied to the datasets prepared for the purpose of further experiments using the rough set approach (see chapter 5.2.5). Therefore, three experiments using IBL1 algorithm were carried out on:

1. random selection of 290 cases from both decision classes
2. set of projections on 9 different condition attributes
3. set of projections on 8 different condition attributes

For each of the new datasets, missing values had been filled in by the most commonly appearing value from the domain of a particular attribute. The modification for handling nominal values was turned off. The k parameter was set to 11. The classification accuracy obtained during a 10-fold cross-validation process was observed. Since only IBL1 algorithm was used, the storage requirements were not observed or compared as they did not change in the experiments.

6.3.1.1. *Experiment 1: selection*

This experiment has been carried out on a dataset in which there were 290 cases from both decision classes. This way we obtained balance in distribution of decision attribute values. In this experiment we aim to observe the changes of classification accuracy obtained for the whole dataset and the one produced through selection. The comparison of results on those two datasets in presented in Table 24.

Table 24. Comparison of classification accuracy obtained on the whole dataset and the dataset after selection

classification accuracy	**whole dataset (867 cases)**	**dataset after selection (580 cases)**
total	64.5% (+-1.34)	65.20% (+-1.65)
CHD=no	96.53% (+-2.24)	62.60% (+-1.63)
CHD=yes	0.69% (+-0.58)	67.80% (+-1.67)

The classification accuracy for the set after selection was a bit higher than for the original dataset. It could be due to the balanced distribution of decision attribute values, but it is also possible that there was less noise or inconsistencies in that randomly selected set than in the original one.

The classification accuracy in CHD=yes class in the dataset after selection reached 67.8% and exceeded the classification accuracy in the other decision class. This is, in general, a desired situation as CHD=yes class is more important from the medical point of view.

6.3.1.2. *Experiment 1: projection to 9 attributes*

Through projection to 9 attributes, 10 different datasets have been obtained and therefore ten rounds of this experiment conducted. In each round we observed what effect on classification accuracy had the elimination of one condition attribute.

The results of this experiment are gathered in Table 25.

Table 25. Results after projection - 9 attributes left

experiment number	number of condition attrib	omitted attrib	classification accuracy
1	9	smoking mother	58.12%
2	9	smoking father	58.26%
3	9	paternal age	59.37%
4	9	birth weight	60.20%
5	9	results of cytogenetic exam	60.28%
6	9	fetal age	60.86%
7	9	maternal age	61.48%
8	9	sex	61.50%
9	9	obstetrical history	62.68%
10	9	place of residence	62.86%

The results show that, elimination of any attribute causes a decrease in the classification accuracy. Elimination of *smoking mother* or *smoking father* attributes has the worst effect on classification accuracy. It drops respectively by 6.38% and 6.24% comparing to the classification accuracy obtained for the dataset containing all 10

condition attributes. The attribute *place of residence* caused the least decrease of the classification accuracy. It is very interesting to observe that, even though the classification accuracy in all of those experiments was lower than the best classification accuracy obtained by IBL1 on the whole dataset, it still exceeded the best classifications obtained on the whole dataset by IBL2 and IBL3 algorithms. Perhaps, it is worth consideration, whether the tradeoff between storage requirements reduction and classification accuracy is better when using IBL2, IBL3 or perhaps IBL1 but with a reduced number of attributes.

The ranking of attributes obtained in this experiment is different from the ranking obtained in an analogical experiment using rough set approach (see chapter 5.2.5.2). In this experiment, the elimination of attribute *smoking mother* causes the greatest reduction in classification accuracy, and therefore it should be the last attribute to be eliminated. In experiment from chapter 5.2.5.2 that elimination of that attribute results in the smallest reduction of quality of approximation. That means that removing this attribute would cause the least harm.

6.3.1.3. *Experiment 3: projection to 8 attributes*

This experiment has been carried out on five datasets prepared for the experiment 3 using rough set approach (see chapter 5.2.5.3). These datasets have been chosen in order to keep the same input in respective experiments using different approaches. The number of condition attributes has been reduced to 8 in each of the prepared datasets. In each dataset we observed the effect of elimination of two particular condition attribute on classification accuracy.

The results of this experiment are presented in Table 26.

Table 26. Results after projection - 8 attributes left

experiment number	number of condition attrib	omitted attrib	classification accuracy
1	8	smoking mother, paternal age	55.41%
2	8	smoking mother, birth weight	53.98%
3	8	smoking father, birth weight	53.67%
4	8	smoking father, smoking mother	52.34%
5	8	smoking father, paternal age	50.89%

With the elimination of two attributes the classification accuracy dropped comparing to the experiments where only one attribute was omitted. All, examined by IBL1 algorithm, sets with 8 condition attributes had lower classification accuracy than the whole set with 10 condition attributes to which IBL2 was applied (then classification accuracy reached 56.27%). In the first experiment, where attributes ***smoking mother*** and ***paternal age*** were eliminated, the classification accuracy was a bit higher than the best classification accuracy obtained on the whole dataset by using IBL3 algorithm (then classification accuracy reached 54.96%). But even with elimination of two attributes, the storage requirements in first situation are larger than in the second.

7. Application of decision tree induction method to knowledge extraction

7.1. Methodological elements of induction of decision trees

A decision tree is a structure that consists of nodes and branches. The starting node is usually referred to as the *root*. Each non-terminal node represents a single test that checks the value of a condition attribute (called *splitting* or *test attribute*) connected with that node. For every possible test result a branch, representing particular value of the attribute, leads to a node on a lower level. In the end, a terminal node, also called a *leaf*, is reached. Each leaf is labelled with one class label representing a given class samples. When a leaf is reached and a decision on the class assignment can be made according to which class has been associated with that leaf [69].

Classification process using decision tree is done in a following manner: Starting from the root, the value of an attribute in the currently checked node is verified. Next, the branch corresponding to a particular value of the attribute leads us to a lower-level node. The process is repeated for a sub-tree associated with a lower-level node until a terminal node, pointing decision class, is reached [31].

All trees can be generally divided into

- binary trees and
- non-binary trees.

In *binary trees* from every non-terminal node there are always two branches leaving it. There is always only one branch entering any node except the root, which is not entered by any branch. For every binary tree there are n terminal nodes and n-1 non-terminal nodes. It is conventional to make true decisions branch right and false branch left.

In case of *non-binary trees*, more than two branches may leave a non-terminal node, but again only one can enter every non-root node. For non-binary trees, a test

performed at a node results in a partition of two or more disjoint sets that all together cover every possibility [69].

Any decision tree can be alternatively represented as a set of rules specifying allocation of cases to decision classes. Each path, in a tree leading to a terminal node corresponds to a decision rule that is a conjunction of various tests. All paths in a decision tree are mutually exclusive, and for any new cases there can only be one path in a tree that shall be satisfied.

7.1.1. *Decision tree induction algorithm*

The process of learning the structure of a decision tree from the data is known as *tree induction.* Most of the tree learning algorithms are based on a heuristic schema of *Top Down Induction of Decision Trees* (TDIDT) presented in Table 27.

Table 27. Basic schema of Top Down Induction of Decision Trees

```
-------------------------------------------------------------------------------------------------------
     function BuildTree
              (Input: S – set of training examples,
                      A – set of condition attributes,
                      SS – split selection method
               Output: a decision tree rooted at node N
               );
     begin
  1  initialize a root node N of the tree
  2  if all cases of TS are of the same decision class C then
  3     return N as the leaf node labeled with the class label C
  4  else
  5     if attribute-list A == {} then
  6        return N as the leaf node labeled with the most common
     class in    TS
  7     else
  8        begin
  9           Apply SS to select best-split attribute from A
  10          Label node N best-split attribute
  11          for each value a_i of the best-split attribute do
  12             begin
  13                Let S_i denote a set of samples in S with
```

```
                    best-split==a_i
14                  Let N_i =
                       BuildTree (S_i, SS, A\{best-split attribute})

15                  Create a branch from N to N_i labeled with the
                       test: best-split = a_i
16               end
17          end
         end { function }
```

It is assumed, that as an input, a set *S* of training examples is available. The construction of the tree starts with making a single node *N* representing the whole training set. If all cases from the *S* are from the same decision class (line 2), then the node *N* becomes a leaf labelled *C* (line 3) and algorithm stops. Otherwise, the set *A* of condition attributes is examined according to the split selection method *SS* in order to select a splitting attribute called *the best split* (lines 9-10). The splitting attribute is used to partition the training set *S* into a set of separate classes $S_1,...S_v$, where S_i, i=1..v contains all those cases from *S* for which *splitting attribute*=a_i (line 13). For each value a_i of the *splitting attribute* a branch labelled V_i is created and to each branch V_i a set S_i of cases is assigned (line 15). The partitioning procedure is repeated recursively for each descendant node to form a decision tree for each partition of cases (line 14). it is important to note, that once an attribute has been chosen as a *splitting attribute* at a given node, it does not have to be considered in any of the descending nodes (line 14).

The main issue in TDIDD algorithm is choosing the *splitting attribute* for building the test, according to which the set of examples in the node will be divided. It is aimed to find such a test which when applied shortens the path leading through the node to the leaves, pointing the decision class [8]. This will be achieved, when in every subset connected to the branches coming out of the node, all or at least majority of examples will be from the same decision class. The choice of test should be made on the grounds of a split selection method which should maximize the accuracy of the constructed decision tree or, in other words, minimize the misclassification rate of the tree. Most of the split selection methods used in such tools as ID3 or C4.5 belong to the class called *impurity-based split selection methods* and find the *splitting attribute* of a node by minimizing an impurity measure such as e.g. the entropy.

7.1.2. *Split selection methods*

One of the most popular split selection method is *information gain* used in algorithm of decision tree induction called *ID3* [49] created by Quinlan. In order to define it, lets introduce *entropy measure*.

Let S be a training set containing examples belonging to one of k decision classes, denoted by $K_1,..., K_k$. Let n be the number of examples in S and n_i be the cardinality of class K_i.

Entropy connected with the classification of set S is defined as:

$$Ent(S) = -\sum_{i=1}^{k} p_i \log_2 p_i$$

where p_i is the probability that a randomly chosen example from S belongs to class K_i, estimated by $\frac{n_i}{n}$.

The smaller the entropy value is, the greater imbalance in set S of distribution of cases between decision classes

Entropy measures expected number of bits needed to code information about randomly chosen example from S [8] and therefore, in the formula above, the bases of logarithm is equal to 2.

In case of a binary classification (i.e., there are only two possible classes and $k=2$) entropy takes value from <0;1>. The maximum value of 1 is reached when $p_1 = p_2 = 0.5$ that is when there is a equal distribution of cases between the two decision classes. The minimal value of entropy is observed when all examples belong to one class.

In situation when an attribute a is used in the test in a node of a decision tree, *conditional entropy* is calculated.

Let attribute a take p different values $\{v_1,..v_p\}$. The test constructed in a tree node by ID3 asks what the value of attribute a is, i.e., a division of set S to subsets $\{S_1,...,S_r\}$. Subset S_j contains examples having value v_j on attribute a (j=1,...,p). Let the cardinality of subset S_j be denoted as n_{S_j}.

Conditional entropy of division of set S into subsets according to attribute a is defined as follows:

$$Ent(S \mid a) = \sum_{j=1}^{p} \frac{n_{S_j}}{n} Ent(S_j).$$

The smaller the value of $Ent(S \mid a)$, the greater homogeneity of classification for the examples divided into subsets.

The *information gain* resulted by using attribute *a* for building the test dividing the training set *S* is defined as:

$$Gain(S, a) = Ent(S) - Ent(S \mid a).$$

$Gain(S, a)$ is the expected reduction of entropy caused by knowing the value of attribute *a*. In other words, it represents the gain of information about classification of examples, when the value of attribute *a* is given.

ID3 algorithm calculates the information gain for every attribute and divides a node according to the attribute with the highest information gain.

It has been observed that information gain puts in favour attributes with large domain. Since it is not a desired feature, an additional measure called *split information* was created in order to judge the division of set of examples in terms of values from the attribute *a*'s domain.

Split information measure is defined as follows:

$$Split(S \mid a) = \sum_{j=1}^{r} \frac{|S_j|}{|S|} \log_2 \frac{|S_j|}{|S|}$$

where S_j is a subset of *S* containing examples characterized by j^{th} value of attribute *a*, *r* is the number of different possible values of attribute *a*.

Split information is used to define a new split selection method for a node of a decision tree called *gain ratio*.

It is defined as follows:

$$Gain\ ratio(S \mid a) = \frac{Gain(S, a)}{Split(S \mid a)}.$$

It is used in a proposed by Quinlan C4.5 [50] algorithm of decision tree induction, which is a successor of ID3. C4.5 chooses the attribute which maximizes the gain ratio as a test in a node but optionally it can also be asked be a user to use the information gain measure.

All the analysis in this work were performed using implementation of the original C4.5 algorithm inducing decision trees proposed by Quinlann. This algorithm has been chosen because it is equipped with approach to dealing with missing values.

7.1.3. *Dealing with missing values*

Real datasets often have missing or unknown values of some attributes. Such is also the analysed dataset of children with congenital heart diseases and Down syndrome. Existence of missing values has influence on the process of induction of the decision tree but also on the process of classification of test or new cases. In literature, many approaches to dealing with missing values in decision trees have been proposed. Some of them simply fill in the missing values by a particular value fro the attributes domain (e.g., the most often occurring or the average value) in the pre-processing phase, but there are also more sophisticated approaches like the one used in C4.5 decision tree induction algorithm.

Let S be the set of cases on the basis of which a splitting attribute for a node is selected. Let a be a potential attribute for the test in a node and let S_0 be the subset of cases from S for which the value of a is not known. Obviously, for cases from S_0 it is not possible to determine the result of the considered test. Therefore, to solve this problem, a modification in form of "penalty" function dependent on relative frequency of unknown values of attribute a has been introduced to information gain measure. The information gain, which is a part of gain ratio split selection method used by C4.5, is calculated on the basis of those cases for which the value of a is not missing (i.e., for cases from subset (S-S_0)) but with respect to the "penalty" function and therefore is defined in the following manner:

$$Gain(S,a) = \frac{|S - S_0|}{|S|} Gain(S - S_0, a)$$

If attribute *a* has been selected to form a test in a node, then a case from the subset S_0 is divided and is "partially" assigned to subsets S_j of cases resembling the results of the test in a created node. During such division, a "partial" case is given a weight which is the probability that attribute *a* will have a particular value. This probability is estimated by the frequency of appearance of different values of the attribute among the set of cases in the node. For example, the weight for subset S_j will be $\frac{|S_j|}{|S - S_0|}$. Divided in such manner, the cases with missing value of attribute *a* are considered in the split selection method for nodes at lower levels, but they are taken for the calculations with the assigned weight.

7.1.4. *Decision tree pruning*

The aim of classification is to find a simple classifier that fits well to the training set and can be well generalized to future, unknown cases. However, during the induction process, a tree might overfit the training set i.e., fit to the training set very well, but loose its ability to classify well new data. In situation of overfitting, the classifiers generalization abilities drop. The process of *pruning of decision tree* addresses the problem of overfitting by removing some branches and nodes of the constructed tree. Two basis approaches to avoid overfitting are distinguished:

- prepruning,
- postpruning.

In the prepruning approach, the decision not to further split the training set at a given node, is made during the phase of tree construction when a chosen measure (e.g., information gain) reaches a given threshold. Upon stopping, the node becomes a leaf labelled with the class to with the majority of cases at the node belong.

In the postpruning approach, some branches and nodes are removed from a tree after its construction phase has been completed. A subtree rooted at a node is replaced with a leaf node labelled with class that was the most frequent one among the former branches. The detailed postpruning method used by C4.5 algorithm is described [30].

7.1.5. *Windowing technique to induction of decision trees*

So far, the proposed methods of decision tree induction assumed having as the input the whole training set. However, C4.5 has an option of creating a decision tree on the basis of only a subset of training examples and then modifying the tree according to results of classification on the remaining (not used for induction) set of examples. This technique is called *windowing*. It follows this basis schema: From the whole training set a subset of chosen cardinality is selected and called a *window*. The tree is induced from this window, and then used to classify the examples that were not part of the window. A certain number of incorrectly classified examples is added to the window and the construction of the decision tree starts again. This whole procedure can be repeated for a certain number of times.

7.2. Application of C4.5 to extraction of knowledge about congenital heart defects in Down syndrome

The analysis of congenital heart disease in Down syndrome has been performed using Quinlan's implementation of C4.5 algorithm of decision tree induction. The effect of different split selection methods (information gain, gain ratio) as well as pruning and windowing techniques on estimated classification accuracy of the created tree was observed. The most informative results are gathered in Table 28.

Table 28. Results obtained from using C4.5

tree	split selection		windowing			before pruning			
	info gain	gain ratio	trials	initial window size	window increment	size	total accuracy [%]	CHD=yes accuracy [%]	CHD=no accuracy [%]
1	yes					125	72.6 (+-1.40)	20.69(+-1.20)	98.90(+-1.91)
2		yes				138	72.0 (+-1.21)	18.71(+-1.19)	98.78(+-1.87)
3	yes		50	20	10	134	72.7 (+-1.61)	20.62(+-1.21)	98.92(+-1.90)
4		yes	50	20	10	149	71.9 (+-1.73)	18.75(+-1.30)	98.70(+-1.88)
5	yes		40	10	10	130	72.1 (+-1.67)	18.70(+-1.29)	98.78(+-2.10)
6		yes	40	10	10	141	71.6 (+-1.57)	18.54(+-1.18)	98.26(+-2.13)

During the first two tests, windowing option was not tested, we only compared the two split selection methods: information gain (info gain) and gain ratio. Later, different experiment with windowing technique have been performed. In the Table 28 we present the four best results of these experiments. The comparison of the induced tree before pruning and after pruning was done in terms of size of the induced tree and its classification accuracy calculated through 10-fold cross-validation. The results show, that no matter what parameters have been taken, the pruning process always made a one-node-tree. This single node had the label of CHD=no class and therefore, the tree after pruning assigned all cases to the decision class without congenital heart disease. In that situation, all cases from CHD=yes were misclassified. This is unacceptable, as correct classification of children with congenital heart disease is especially important for doctors. Therefore, we take under consideration only the trees before pruning. They are much bigger as they consist of over 100 nodes, but their classification accuracy is over 70%. This accuracy seems a very high score comparing to results obtained by other analyzed methods, but the complexity of knowledge representation in form of such a tree is also very high. The tree of over 100 nodes is too big and too complex to easily used by medical experts. There would be a few dozens of rules if we transformed the tree to a decision rule set and again. Many of them would be of weak strength and confidence. Thus, even though the obtained classification accuracy is high, those trees were not a welcomed form for medical experts.

Using the windowing technique, did not any bring better results. The number of trials (iterations), initial window size and window increment size has been changed in many ways in order to obtain as good results as possible. The best results are presented in Table 28., but again they are characterized by large tree size, even though the classification accuracy is good (exceeds 70%).

It is interesting to note, that in analogical experiments (1&2; 3&4; 5&6), the ones where information gain was used as the split selection method had a bit better classification accuracy and a bit smaller tree size, comparing to the experiments using gain ratio. Since, none of the attribute domains is continuous, and their cardinalities are rather similar, then the measure of information gain is enough and there is no need to introduce gain ratio.

7.2.1. *Further experiments*

Decision tree induction was also additionally performed on the datasets prepared for the purpose of further experiments using the rough set approach (see chapter 5.2.5). Therefore, three experiments using C4.5 algorithm were carried out on:

1. random selection of 290 cases from both decision classes
2. set of projections on 9 different condition attributes
3. set of projections on 8 different condition attributes

The split selection criterion remain unchanged in all experiments and it is information gain. We observe changes in tree size and classification accuracy measured during 10-fold cross-validation.

7.2.1.1. *Experiment 1: selection*

This experiment has been carried out on a dataset in which the number of cases from both decision classes was equal i.e., there was a balance in distribution of decision attribute values. The results of applying decision tree approach to this dataset are presented in Table 29.

Table 29. Comparison of results obtained on the whole dataset and the dataset after selection

set	before pruning				after pruning			
	size	total accuracy [%]	CHD=yes accuracy [%]	CHD=no accuracy [%]	size	total accuracy [%]	CHD=yes accuracy [%]	CHD=no accuracy [%]
whole set	125	72.60 (+-1.40)	20.69 (+-1.20)	98.90 (+-1.91)	1	66.60 (+-0.10)	0.00 (+-0.10)	100.00 (+-0.10)
set after selection	134	68.60 (+-1.39)	69.20 (+-2.01)	68.28 (+-1.83)	32	62.60 (+-2.03)	67.24 (+-1.99)	57.93 (+-1.98)

For the tree before pruning, the tree induced from the whole dataset is smaller and has a better classification accuracy.

However, the pruned tree induced from the set after selection is much better than the tree created from the whole dataset. First look at the results in Table 29, could be misleading as the tree from the whole set is smaller in size and has higher total classification accuracy, but comparison of the classification accuracies in each class prove, that the tree from the smaller set is better. The pruned tree created from the whole set has just one node and classifies all cases to CHD=no class. This way, all important cases with congenital heart disease are misclassified. The pruned tree induced from the dataset with balanced classes is a 32-node tree. This size is still acceptable for medical experts. Moreover, this tree classifies correctly over 67% of cases from CHD=yes class. The classification accuracy in the class with congenital heart disease is higher than in class without it, which is a very desired situation, from medical point of view.

7.2.1.2. *Experiment 2: projection to 9 attributes*

Through projection to 9 attributes, 10 different datasets have been obtained. For each of those sets we observed the effect of elimination of one condition attribute on classification accuracy.

The results of this experiment are gathered in Table 30.

Table 30. Results after projection - 9 attributes left

tree	number of condition attributes	omitted attribute	before pruning		after pruning	
			size	accuracy [%]	size	accuracy [%]
1	9	smoking father	125	71.0	1	66.6
2	9	smoking mother	123	71.6	1	66.6
3	9	results of cytogenetic exam	139	71.6	1	66.6
4	9	sex	97	69.9	1	66.6
5	9	fetal age	91	70.2	1	66.6
6	9	birth weight	89	69.8	1	66.6
7	9	paternal age	115	70.5	1	66.6
8	9	maternal age	114	72.1	1	66.6
9	9	obstetrical history	97	70.7	1	66.6
10	9	place of residence	56	68.5	1	66.6

Each of the 10 pruned trees is actually a one-node tree which classifies all cases to CHD=no class. It is unacceptable from the medical point of view, therefore we shall turn to trees before pruning. Elimination of attribute *maternal age* or *smoking mother* or *results of cytogenetic examination* has the smallest decreasing effect on classification and therefore, these attributes could be dropped as the first ones. *Place of residence* should be at the end of the list of candidates for elimination as its elimination from the dataset causes the biggest reduction in classification accuracy. Again, we can observe that the "ranking" of attributes obtained using decision trees differs from the rankings obtained by other approaches in previous Sections. It is due to methodological difference between approaches.

7.2.1.3. *Experiment 3: projection to 8 attributes*

This experiment has been carried out on five datasets prepared for the experiment 3 using rough set approach (see chapter 5.2.5.3). These datasets have been chosen in order to keep the same input in respective experiments using different approaches. We aim at observing the changes in classification accuracy for different sets with 8 condition attributes. The results of this experiment are gathered in Table 31.

Table 31. Results after projection - 8 attributes left

tree	number of condition attribute	omitted attribute	before pruning		after pruning	
			size	accuracy [%]	size	accuracy [%]
1	8	smoking mother, smoking father	123	71.0	1	66.6
2	8	smoking father, paternal age	107	70.2	1	66.6
3	8	smoking mother, birth weight	89	69.9	1	66.6
4	8	smoking father, birth weight	79	69.4	1	66.6
5	8	smoking mother, paternal age	111	70.5	1	66.6

Again, we face the problem of one-node pruned trees, which are not acceptable for medical experts. About trees before pruning, it can observed that their classification accuracy is still quite high despite elimination of two attributes. The tree induced from dataset with omitted *smoking father* and *smoking mother* has almost the same size and accuracy as trees from sets where only one of those attributes was eliminated. Those

two attributes, in that situation, can be seen as rather uninformative. Moreover, the results show that, from the examined pairs of attributes, the pair ***smoking father, birth weight*** brings the biggest reduction in classification accuracy, when eliminated from the dataset.

8. Application of logistic regression to knowledge extraction

8.1. Methodological elements of logistic regression

Logistic regression is a linear statistical method for classification [23]. It is aimed at modeling the posterior probabilities of decision classes via function in condition attributes. To describe this approach more precisely, let us introduce the following notation:

A set of condition attributes is referred to as an *input variable* and denoted by the symbol X. If X is a vector, its components can be accessed by subscripts X_j. A set of decision attributes (in our research there is only one decision attribute) is referred to as an *output variable* and denoted by G (for group). Observed values of variables are written in lowercase; hence the i^{th} observed value of X is written x_i (where x_i is again a scalar or vector). Thus, the training set is composed of measurements (x_i, g_i). The task is to make a good prediction, of the output G, denoted by $\hat{G}$ on the basis of the value of an input vector X.

In general, the logistic regression model has the form:

$$\ln \frac{\Pr(G=1|X=x)}{\Pr(G=K|X=x)} = b_{10} + b_1^T x$$

$$\ln \frac{\Pr(G=2|X=x)}{\Pr(G=K|X=x)} = b_{20} + b_2^T x$$

$$\ln \frac{\Pr(G=K-1|X=x)}{\Pr(G=K|X=x)} = b_{(K-1)0} + b_{K-1}^T x$$

where $\Pr(G=j|X=x)$ is the probability that an input cases x will be classified to class j, and b_{i0} is an interceptor, and b_i are coefficients.

As seen the model is specified in terms of *logit* or *logistic transformations* which are generally defined as:

$$\ln\{\frac{p}{1-p}\},$$

where p is a probability value.

Note, that the value of this logarithm can theoretically assume any value between minus and plus infinity.

From the above model results that

$$\Pr(G = k \mid X = x) = \frac{\exp(b_{k0} = b_k^T x)}{1 + \sum_{l=1}^{K-1} \exp(b_{l0} + b_l^T x)}, k = 1,..., K-1,$$

$$\Pr(G = K \mid X = x) = \frac{1}{1 + \sum_{l=1}^{K-1} \exp(b_{l0} + b_l^T x)}.$$

For the sake of notation simplicity we denote the probabilities $\Pr(G = k \mid X = x) = p_k(x;q)$

where $q = \{b_{10}, b_1,..., b_{(K-1)0}, b_{K-1}\}$.

When there are only two classes, $K = 2$, this model is especially simple and takes the form:

$$\ln \frac{\Pr(G = 1 \mid X = x)}{\Pr(G = 2 \mid X = x)} = b_0 + b^T x$$

Thus,

$$\Pr(G = 1 \mid X = x) = \frac{\exp(b_0 = b^T x)}{1 = \exp(b_{l0} + b^T x)},$$

$$\Pr(G = 2 \mid X = x) = \frac{1}{1 + \exp(b_{l0} + b^T x)}.$$

Logistic regression models are fit by maximum likelihood, using the conditional likelihood of G given X. Since $\Pr(G \mid X)$ completely specifies the conditional distribution, the *multinominal* distribution is appropriate. The log-likelihood for N observations is

$$\mathbf{l}(q) = \sum_{i=1}^{N} \log p_{g_i}(x_i;q)$$

where $p_k(x_i;q) = \Pr(G = k | X = x_i;q)$.

We discuss in detail the two-class case, since the algorithms simplify considerably. It is convenient to code the two-class g_i via a 0/1 response y_i, where $y_i = 1$ when $g_i = 1$ and

$y_i = 0$ when $g_i = 2$. Let $p_1(x_i;q) = p(x;q)$ and $p_2(x;q) = 1 - p(x;q)$. The log-likelihood can be written

$$\begin{aligned}\mathbf{l}(b) &= \sum_{i=1}^{N} \{y_i \log p(x_i;b) + (1-y_i)\log(1-p(x_i;b))\} \\ &= \sum_{i=1}^{N} \{y_i b^T x_i - \log(1+e^{b^T x_i})\}\end{aligned}$$

Here $b = \{b_{10}, b_1\}$, and we assume that the vector of inputs x_i includes the constant term 1 accommodate the intercept.

To maximize the log-likelihood, we set its derivatives to zero. These *score* equations are

$$\frac{\partial \mathbf{l}(b)}{\partial b} = \sum_{i=1}^{N} x_i (y_i - p(x_i;b)) = 0,$$

which are $p + 1$ equations *nonlinear* in b. Notice that since the first component of x_i is 1, the first score equation specifies that $\sum_{i=1}^{N} y_i = \sum_{i=1}^{N} p(x_i;b)$; the *expected* number of class ones matches the observed number (and hence also class twos).

Solving this equation using, for example, the Newton-Raphson algorithm we find iteratively required constant b.

Logistic regression models are used mostly as a data analysis and inference tool, where the goal is to understand the role of the input variables in *explaining* the outcome. Typically many models are fit in a search for a parsimonious model involving a subset of the variables, possibly with some interactions terms.

In practice, for a two-class model the values of intercept and coefficients b are in fact the values of appropriate statistics. Thus, it is necessary to verify statistic null hypothesis that the intercept or coefficient in question is zero, while all the others are

not (in the tool (STATISTICA 6.0) we used for the calculations, it is done using Wald's test [60]). To simplify testing of this hypothesis so called *p-level* is calculated for the interceptor and each coefficient. The *p*-level represents the probability of erroneous rejection of the null hypothesis. Thus, the *p*-level represents the probability of error that is involved in accepting our observed result as valid, that is, as "representative of the population". The higher the p-level, the less we can believe that the observed relation between variables in the sample set is a reliable indicator of the relation between the respective variables in the population. In many areas of research, the *p*-level of 0.05 is treated as a "border-line acceptable" error level, and so was in this thesis. Thus, any b will be treated as statistically insignificant if its *p*-level was greater than 0.05 i.e., there were no grounds to reject the null hypothesis.

The above mentioned two-sided test of hypothesis is very close to inference based on *confidence intervals*.

Let L and U be two statistics of the training set such that $L \leq U$. The interval $<L,U>$ such that

$$\Pr(L \leq b \geq U) \geq 1-\alpha,$$

where α is a given probability,

is called $(1-\alpha)100\%$ confidence interval for b parameter. The probability $1-\alpha$ is called confidence level for the $<L,U>$ interval.

It means that the confidence interval includes the estimated parameter b with the probability $1-\alpha$.

8.2. Application of logistic regression to extraction of knowledge about congenital heart defects in Down syndrome

The analysis of congenital heart disease in Down syndrome has been performed using Statistica 6.0 -a tool for statistical analysis. Missing values have been filled in by the most commonly appearing value from the particular attribute domain.

In order to perform logistic regression calculations, nominal attributes have been recoded to numerical. Each of the possible values from particular attribute's domain has been given the following natural number starting from 1. Since most attributes had a two-value domain, the given numbers were 1 and 2. Table 32 gives few details about the values of the attributes after transformation. Attributes *place_of_residence* and *cytogenetic_exam* had the largest domains and therefore, they have the two highest maximum values.

Table 32. Means, standard deviations, minimum and maximum values of original attributes transformed to numerical attributes

variable	mean	st. dev.	minimum	maximum
place_of_residence	2,119898	1,093771	1,000000	4,000000
sex	1,445153	0,497300	1,000000	2,000000
cytogenetic_exam	1,979592	0,251873	1,000000	3,000000
fetal_age	1,334184	0,472005	1,000000	2,000000
birth_weight	1,184949	0,388504	1,000000	2,000000
maternal_age	1,322704	0,467809	1,000000	2,000000
paternal_age	1,510204	0,500215	1,000000	2,000000
obstetrical_hist	1,243622	0,429542	1,000000	2,000000
smoking_father	1,036990	0,188857	1,000000	2,000000
smoking_mother	1,020408	0,141482	1,000000	2,000000
CHD	0,655612	0,475472	0,000000	1,000000

The results of performing logistic regression on the dataset of children with congenital hear disease and Down syndrome are gathered in Table 33. In column *Estimate* are presented the calculated values of intercept and coefficients, column *Wald Stat.* presents the values of Wald statistic calculated during the Wald test and the column *p* shows the *p-level* corresponding to particular value of the Wald statistic. Marked red are the results for the attribute for which the *p-level* was below the border

value of 0.05. For that attribute, there grounds to reject the null hypothesis and therefore, we consider the coefficients for that attribute as statistically significant.

Table 33. Results of logistic regression on the whole dataset

Effect	Column	Estimate	Standard Error	Wald Stat.	p
Interc	1	1,704252	79,55497	0,000459	0,982909
place_of_residence	2	0,128444	0,06697	3,678104	0,055132
sex	3	0,384108	0,14708	6,820150	0,009013
cytogenetic_exam	4	-0,309294	0,28401	1,185978	0,276142
fetal_age	5	0,153206	0,16712	0,840421	0,359277
birth_weight	6	0,143899	0,19532	0,542787	0,461280
maternal_age	7	-0,270047	0,19651	1,888530	0,169368
paternal_age	8	-0,127710	0,18163	0,494390	0,481976
obstetrical_hist	9	-0,044931	0,17590	0,065247	0,798387
smoking_father	10	-0,281156	0,41034	0,469463	0,493234
smoking_mother	11	-0,197279	0,60508	0,106300	0,744396
Scale		1,000000	0,00000		

The results show, that only coefficients for attribute *sex* are statistically significant. In particular, the *p-level* of the intercept reached 0.98 which indicates that there is a 98% probability that the relation between variables found in our dataset is a "fluke".

Table 34 presents confidence intervals of the Estimates. The wider the interval is, the less precise we can be about results of the attribute. The best, in terms of having the smallest interval, is attribute *sex* and the worst results were obtained for the intercept.

Table 34. Confidence intervals of Estimates for the whole dataset

Effect	Column	Lower CL 95, %	Upper CL 95, %
Interc	1	-154,221	157,6291
place_of_residence	2	-0,003	0,2597
sex	3	0,096	0,6724
cytogenetic_exam	4	-0,866	0,2474
fetal_age	5	-0,174	0,4808
birth_weight	6	-0,239	0,5267
maternal_age	7	-0,655	0,1151
paternal_age	8	-0,484	0,2283
obstetrical_hist	9	-0,390	0,2998
smoking_father	10	-1,085	0,5231
smoking_mother	11	-1,383	0,9887

All those results, prove that the analyzed dataset cannot bring satisfactory results in terms of classification accuracy. The classification accuracy was calculated through 10-fold cross-validation. For every case from the test set, the posterior probability of belonging to decision class with congenital heart disease was calculated. An error (misclassification) occurred when the probability was lower than 0.5 while the case was actually classified as belonging to CHD=no class or when it was equal or above 0.5 while the case was actually classified as belonging to CHD=yes class. We have obtained classification accuracy of 67,24% , which is not high, but still the highest from those obtained by other analyzed methods. The confusion matrix is shown in Table 35. The number of misclassified cases from the CHD=yes class is sadly high. The accuracy for this class is only 3,79%. From medical point of view, accuracy in classification on CHD=yes class is much more important the other class. However, it must be admitted that the number of correctly classified cases with congenital heart disease is a bit higher than in other analyzed methods.

Table 35. Confusion matrix for the whole dataset

Confusion Matrix (sum over 10 passes)		
	PREDICTED	
ACTUAL	CHD=no	CHD=yes
CHD=no	572	5
CHD=yes	279	11

Average Accuracy [%]		
	Correct	Incorrect
Total	67.24 (+-1.24)	32.76 (+-1.24)
CHD=no	99.13 (+-1.43)	0.87 (+-1.43)
CHD=yes	3.79 (+-1.01)	96.21 (+-1.01)

The graphical illustration of accuracy of classification as well as the confusion matrix is presented in Figure 16.

It can be observed that for most cases from the analyzed set the calculated probability was lower than 0.5. Therefore, all most all cases from the CHD=no class (marked as 0 on the observed values ax) were classified correctly, and most all the cases from

CHD=yes class (marked as 1 on the observed values ax) were misclassified. The fact that 66% of all cases belong to the class without congenital heart disease, had a strong effect on the value of classification accuracy.

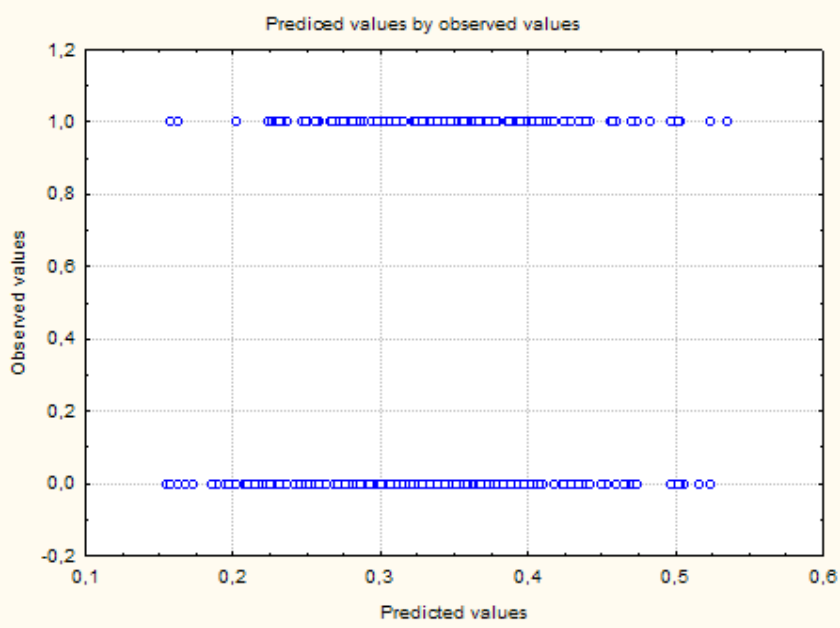

Figure 16. Predicted values by observed values for the whole dataset

8.2.1. *Further experiments*

Logistic regression was also additionally performed on the datasets prepared for the purpose of further experiments using the rough set approach (see chapter 5.2.5). Therefore, three experiments using logistic regression were carried out on:

1. random selection of 290 cases from both decision classes
2. set of projections on 9 different condition attributes
3. set of projections on 8 different condition attributes

8.2.1.1. *Experiment 1: selection*

This experiment has been carried out on a dataset in which the number of cases from both decision classes was equal. In the original dataset the decision classes were imbalanced, favoring the CHD=no class to which belonged 66% of all cases. The balance was obtained by random selection of 290 cases from the CHD=no class and

adding all (i.e., 290) cases from the CHD=yes class. The results of applying logistic regression to this dataset are presented in Table 36 and Table 37.

Table 36. Results of logistic regression for the dataset after selection

Effect	Column	Estimate	Standard Error	Wald Stat.	p
Interc	1	-178,368	90,97026	3,844467	0,049910
place_of_residence	2	0,045	0,08476	0,279100	0,597292
sex	3	0,449	0,17192	6,834971	0,008939
cytogenetic_exam	4	0,526	0,33692	2,438101	0,118419
birth_weight	5	-0,052	0,19172	0,072345	0,787953
fetal_age	6	-0,364	0,23307	2,435208	0,118638
maternal_age	7	0,219	0,22615	0,941405	0,331917
paternal_age	8	0,332	0,21231	2,443234	0,118033
obstetrical_hist	9	-0,069	0,20662	0,112269	0,737576
smoking_father	10	0,668	0,44082	2,297339	0,129596
smoking_mother	11	0,552	0,63525	0,754831	0,384951
Scale		1,000	0,00000		

Table 37. Confidence intervals of Estimates for the dataset after selection

Effect	Column	Lower CL 95, %	Upper CL 95, %
Interc	1	-356,667	-0,069808
place_of_residence	2	-0,121	0,210917
sex	3	0,113	0,786405
cytogenetic_exam	4	-0,134	1,186440
birth_weight	5	-0,427	0,324191
fetal_age	6	-0,821	0,093099
maternal_age	7	-0,224	0,662657
paternal_age	8	-0,084	0,747971
obstetrical_hist	9	-0,474	0,335738
smoking_father	10	-0,196	1,532145
smoking_mother	11	-0,693	1,796969

The coefficients for attribute *sex* and the intercept are statistically significant (marked red in Table 36) and they have the best (i.e., the smallest) confidence intervals presented in Table 37. The results have improved comparing to the results obtained o the whole dataset where the intercept was not statistically significant.

However, the classification accuracy reached 61,2% and was lower than accuracy observed for the whole dataset. The positive aspect, though, is that, as shown in Table 38,the number of correctly classified cases with congenital heart disease

increased up to 178 cases which is 61,4%. Note, that for the whole dataset, the accuracy on CHD=yes class was only 3,79%.

Table 38. Confusion matrix for the selected dataset

Confusion Matrix (sum over 10 passes)		
	PREDICTED	
ACTUAL	CHD=no	CHD=yes
CHD=no	177	113
CHD=yes	112	178

Average Accuracy [%]		
	Correct	Incorrect
Total	61.20 (+-1.17)	38.80 (+-1.17)
CHD=no	61.03 (+-1.20)	38.97 (+-1.20)
CHD=yes	61.38 (+-1.98)	38.62 (+-1.98)

Figure 17 can be treated as an illustration to the confusion matrix in Table 38. It shows that many cases from the CHD=no class (marked as 0 on the observed values ax) were misclassified because the probability calculated for them exceeded 0.5, but also many cases from CHD=yes class (marked as 1 on the observed values ax) were incorrectly classified because the probability calculated for them was lower than 0.5.

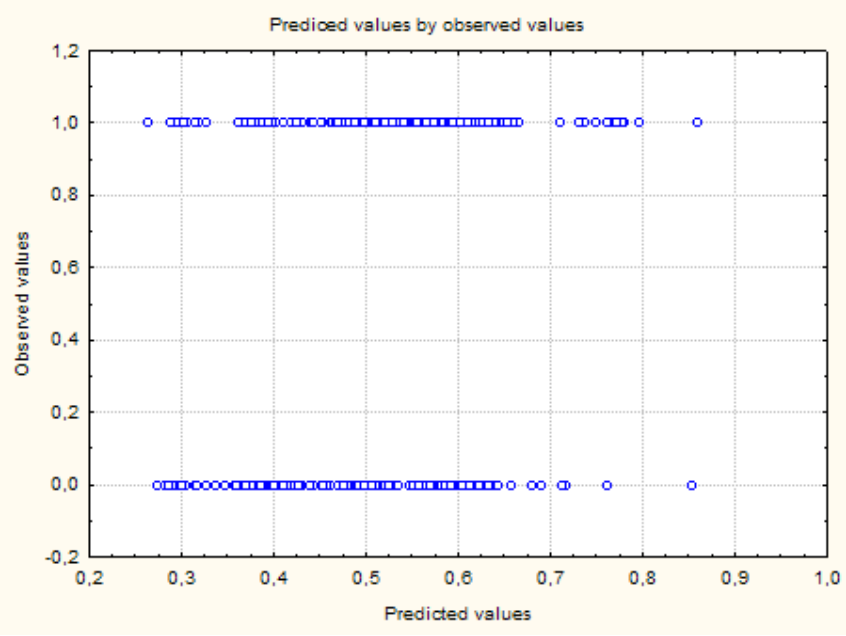

Figure 17. Predicted by observed values for the dataset after selection

8.2.1.2. *Experiment 2: projection to 9 attributes*

Through projection to 9 attributes, 10 different datasets have been obtained and therefore ten rounds of experiment conducted. In each round we observed the effect of elimination of one condition attribute on classification accuracy.

The results of this experiment are gathered in Table 39.

Table 39. Results after projection - 9 attributes left

experiment number	number of condition attrib	omitted attrib	classification accuracy	number of estimates statistically significant
1	9	smoking mother	65.74%	1 (sex)
2	9	smoking father	65.74%	1 (sex)
3	9	birth weight	65.86%	2 (sex, place of residence)
4	9	obstetrical history	65.86%	2 (sex, maternal age)
5	9	fetal age	65.86%	1 (sex)
6	9	maternal age	65.86%	1 (sex)
7	9	paternal age	65.97%	1 (sex)
8	9	sex	66.44%	0
9	9	results of cytogenetic exam	66.09%	1 (sex)
10	9	place of residence	66.09%	1 (sex)

For all of those sets, all cases from decision class CHD=yes were misclassified. Elimination from the original dataset attributes *paternal_age* and in the next test *birth_weight* had the effect on number of statistically significant estimates (the number

rose up to two). The smallest reduction in classification accuracy was observed after elimination of attribute *sex*. Surprisingly, the attributes that had the least effect on quality of approximation in rough set approach, in logistic regression bring the biggest decrease in classification accuracy. Therefore, the ranking of attributes when talking about logistic regression, would be as the Table 39 is sorted.

8.2.1.3. *Experiment 3: projection to 8 attributes*

This experiment has been carried out on five datasets prepared for the experiment 3 using rough set approach. These datasets have been chosen in order to keep the same input in respective experiments using different approaches. In each dataset there are 867 cases described by 8 condition attributes. In each dataset we observed the effect of elimination of two particular condition attribute on classification accuracy. The results of this experiment are gathered in Table 40.

Table 40. Results after projection - 8 attributes left

experiment number	number of condition attrib	omitted attrib	classification accuracy	number of estimates statistically significant
1	8	smoking mother, birth weight	65.74%	2 (sex, place of residence)
2	8	smoking mother, paternal age	65.97%	2 (sex, maternal age)
3	8	smoking father, birth weight	65.86%	2 (sex, place of residence)
4	8	smoking father, smoking mother	65.86%	2 (sex, maternal age)
5	8	smoking father, paternal age	65.97%	1(sex)

It is interesting to observe that four situations elimination of two attributes caused increase in the number of statistically significant estimates. The classification accuracy did not drop much comparing to results obtained on 9-condition-attribute-sets in the previous experiment. However, elimination of an attribute brings reduction to the set size. The smaller, the set is, the faster go the computations, although, the analyzed dataset was still too small to observe reduction in processing time.

9. Comparison of results obtained using rough set theory, IBL, C4.5 and logistic regression

9.1. Obtained classification accuracies comparison

In Table 41 classification accuracies obtained by application of different approaches to the dataset of children with congenital heart diseases and Down syndrome are gathered. Each of those approaches has been applied to the analyzed dataset a couple of times with different parameters and the accuracies presented in Table 41 are the best results obtained for those approaches.

Table 41. Classification accuracies obtained using different approaches

classification accuracy [%]	rough sets	IBL1	C4.5 before pruning	C4.5 after pruning	logistic regression
total	65.64 (+-6.41)	64.50 (+-1.34)	72.60 (+- 1.40)	66.60 (+-0.01)	67.24 (+-1.24)
CHD=no	98.18 (+-2.24)	96.53 (+-2.24)	98.90 (+-1.91)	100.00 (+-0.01)	99.13 (+-1.43)
CHD=yes	1.27 (+-1.08)	0.69 (+-0.58)	20.69 (+-1.20)	0.00 (+-0.01)	3.79 (+-1.01)

The classification accuracy varies from 65.50% to 72.6%. It reached the highest value for unpruned decision tree induced by C4.5 algorithm, however it should be stressed that the complexity of this tree decreases very much its usefulness. But comparing the classification accuracies, one should take under consideration also the percentage of correctly classified cases from the decision class with congenital heart malformation as it is a more important class from medical point of view. The logistic regression approach had the best classification accuracy in class CHD=yes, which reached 3.79%, We shall not consider the 20.69% accuracy in this class for decision tree before pruning as such tree is too complex to use easily. The classification accuracy in

CHD=yes class for a decision tree after pruning turned out to be a total disaster and makes this tree completely useless. The pruned tree consists of only one node and classifies all objects to CHD=no class and therefore, it does not classify correctly even a single object with congenital heart defect. Also for the IBL1 algorithm, the classification accuracies in CHD=yes class is unacceptable low. It is a bit higher for the rough set approach.

Such results of classification accuracy in CHD=yes class are due to too many inconsistencies among the dataset. It is a feature of the dataset itself, independent of the used method.

9.2. Advantages and disadvantages of knowledge form representation in different approaches

9.2.1. *Rough sets*

The rough set approach induces a set of decision rules. This knowledge representation is very easy to understand and use. Many medical experts consider decision rules as their favorite knowledge representation. The big advantage of rough set approach is that it gives both certain and uncertain rules.

One of the disadvantages of decision rules representation, in general, is that its usefulness decreases when the rules are too long or when the set contains too many rules. The rough set approach solves this problem by introducing a *rule length* parameter that can limit the length of induced rules, and *strength* which can limit the set of rules only to rules that are characterized by the strength above that given level.

9.2.2. *Instance Based Learning - IBL 1-3*

There is no explicit knowledge representation form for IBL 1-3 approaches. These algorithms are like "black boxes" that give only the classification decision based on concept description and calculated similarities. Definition of the similarity measure is, moreover, arbitrary to a large extent. One cannot see the relations and dependencies between attributes.

9.2.3. *Decision tree induction - C4.5*

C4.5 algorithm induces knowledge in a form of decision tree. It is a very intuitive representation form. Its hierarchical form can be a guidance for medical experts on which questions and in which order to pose to the patient. Moreover, the decision trees have the advantage that the induced from them knowledge can be alternatively presented in form of decision rules. Each paths in the tree can become one decision rule. However, a set of such rules should also be checked in order to verify if there are no redundant rules.

One disadvantage of tree representation is that the more complex the tree is, the more difficult to use it becomes. Tree complexity is also connected with the problem of overfitting. However, these problems can be addressed by tree pruning. Unfortunately, for the analyzed dataset, the decision tree after pruning is completely useless as it misclassifies all the objects from the medically more important decision class (i.e., CHD=yes class).

9.2.4. *Logistic regression*

This form of knowledge representation is very interesting and not very commonly used. The knowledge is represented in form of posterior probabilities of decision classes modeled via linear function in condition attributes. This way we not only receive classification decision, but also find out what the probabilities of classifying the analyzed case to each class are. The form of the regression function hides, however, the actual influence of the particular attributes and subsets of these attributes on final decision.

10. Conclusions and final remarks

In this thesis a dataset from the database of the Polish Registry of Congenital Malformations has been dealt with. A part of the Registry concerned with children with Down syndrome became the special interest of the analysis. The chosen dataset describes 867 children with Down syndrome, among which 290 also suffer from congenital heart defect. Each object from the dataset is described by 10 condition attributes and one decision attribute, telling whether the child does or does not have a congenital heart defect. This dataset has been gathered by many different physicians treating children with Down syndrome and congenital heart defects, however, without any control of the quality of the data. This might be a reason why the dataset contains both missing attribute values and many inconsistencies. Some preprocessing techniques like discretization, elimination of duplicates, etc., have been applied to the dataset, which do not decrease, however, the data inconsistency. Then, an attempt to extract knowledge about existence of congenital heart defect among children with Down syndrome from this dataset has been made. The following approaches to knowledge extraction have been applied to the dataset:

- rough set theory,
- instance based learning,
- decision trees induction,
- logistic regression.

The average classification accuracy obtained by different approaches varies from 65.5% to 72.6%. The highest value was observed for a decision tree before pruning. It needs to be noted, however, that the complexity of this tree has a negative effect on its usefulness and therefore, it had been excluded from further analysis. The lowest classification accuracy was obtained by instance based learning approach. The classification accuracy for particular classes ranges from 96.5-99.13% for the class of children without congenital heart defect and from 0-3.79% for the class of children with congenital heart defect. The low classification accuracy for the decision class with congenital heart defect brings us to the conclusion that the analyzed dataset contains too many inconsistencies. Let us stress that this is a feature of the dataset, and not of the data analysis method being used.

Such a large amount of inconsistencies appearing in the analyzed dataset might be due to the fact that there is no quality control over data at the phase of registration of a new case of congenital malformation. The paper registration forms filled in by physicians all over the country sometimes contain information wrongly classified, misinterpreted or omitted. Introducing a registration through web side, could be a solution to this problem. Switching to a computer solution at all data gathering phases would have a positive effect on data quality and would surely eliminate many inconsistencies.

Moreover, the inconsistencies in the dataset might have been caused by the fact that there were too few condition attributes considered in the analysis. Perhaps, some attributes which distinguish well the decision classes, were not taken under consideration or are not available in the database of the Polish Registry of Congenital Malformations. Therefore, a thorough extension of gathered information about children with congenital malformations should be considered.

Putting aside the question of the quality of data, an important aim of the thesis was a comparison of the different approaches to knowledge extraction from the data. The rough set approach induces knowledge in form of a set of decision rules. This representation is very legible for showing relations between condition attributes or groups of attributes and the decision attribute. The distinction between certain and uncertain rules is a very valuable aspects of this approach. Moreover, such parameters as rule strength or confidence are informative parameters that show the value of each rule and can be also used to manipulate the size of the decision rule set. A very important advantage of the rough set approach is that it points out existing in the analyzed dataset reducts i.e., minimal subsets of attributes ensuring the same quality of classification as the entire set. The idea of reducts is very interesting as it allows identification of redundant attributes. By eliminating them, a dataset size reduction can be made. Moreover, the rough set approach shows an intersection of all reducts in the information system called a *core*. This allows us to identify the subset of attributes that is absolutely necessary to preserve the classification accuracy obtained on the whole set of attributes.

The instance based learning approach does not have an explicit form of knowledge representation, which might be considered as a disadvantage as relations and dependencies between attributes are not seen. Moreover, the definition of the similarity

measure is, arbitrary to a large extent. The fact that the instance based learning is an incremental approach might be considered as its advantage as the classifier is modified whenever a new training objects arrives, without the need to start the construction of the classifier from the scratch. However, this also makes the instance based learning approach vulnerable to the order of appearance of the objects from the training set. All in all, the instance based approach is commonly used due to its simplicity and the fact that its results of classification accuracy are comparable to results of other methods.

The decision tree induction approach is characterized by a very intuitive form of knowledge representation. Its hierarchism shows the importance of different attributes at different stages of classification. Decision trees can adjust well to the training set and give very good classification accuracies on the training set. This is, however, very often connected with a big complexity of its structure and a possible danger of overfitting. Therefore, decision trees undergo the process of pruning. The pruned tree is meant to be of smaller size and not significantly lower classification accuracy. However, the process of pruning might end with loss of valuable information.

The logistic regression is a statistical approach. It has an interesting but not as intuitive and understandable as decision rules or trees, form of knowledge representation. The knowledge is represented in a form of posterior probabilities of decision classes. Thus, we find out what the probabilities of classifying an analyzed object to each of the decision classes are. This is an additional and very interesting information apart from receiving the classification decision itself. The disadvantage of the logistic regression approach is that this form of the regression function does not show the actual influence of the particular attributes and groups of attributes on the final classification decision.

The problem considered in this thesis could also be investigated further along the following possible lines:

- extension of the analysis to a dataset with more decision classes, indicating not only the presence or absence of the congenital heart defect, but telling also what particular kind of a heart defect it is;
- knowledge extraction from other parts of the Polish Registry of Congenital Malformations;

- application of other knowledge extraction approaches, e.g., neural networks or approaches taking into account preference order in domains of condition and decision attributes;
- analysis of the performance and scalability of applied knowledge extraction methods.

11. References

[1] Aha D., Tolarating noisy, irrelevant and novel attributes in instance-based learning algorithms, *Int. J. Man-Machine Studies*, 36 (1992 267-287.

[2] Aha D.W., Case-Based Learning Algorithms. *Proceedings of the Case-Based Reasoning Workshop*, May 1991, Morgan-Kaufmann, pp. 147-158.

[3] Aha D.W., Kibler D., Albert M.K., Instance-based learning algorithms, *Machine learning* 6 (1991) 37-66.

[4] Barlow G.M., Chen X.N., Shi Z.Y. et al, Down syndrome congenital heart disease: a narrowed region and a candidate gene, *Genet-Med* Mar-Apr; 3 (2) (2001) 91-101.

[5] Benke P.J., Carver V., Donahue R., Risk and Recurrence Risk of Down Syndrome, www.nas.com/downsyn/benke.html

[6] Brandt S., Analiza danych - metody statystyczne i obliczeniowe, PWN, Warszawa 1998.

[7] Chmielewski M.R., Grzymala-Busse J.W., Global discretization of continuous attributes as preprocessing or machine learning, in: [36], pp.294-297.

[8] Cichosz P., *Systemy uczące się*, WNT, Warszawa 2000.

[9] De la Rochebrochard E., Thonneau P., Paternal age and maternal age are risk factors for miscarriage; results of a multicentre European study. *Human reproduction* 17 (6) (2002) 1649-1656.

[10] De-Rubens Figueroa J., Del-Pozzo-Magana B et al., Heart malformations in children with Down syndrome, *Revista Espanola de* cardiologia, Sept; 56 (9) (2003) 894-899.

[11] Digilio M.C., Marino B., Canepa S.A. et al, Congenital heart defect in sibs with discordant karyotype, *American Journal of Medical Genetics*, Nov 2;80(2) (1998) 169-72.

[12] Digilio M.C., Marino B., Genetic predisposition to ventricular septal defect in Down syndrome, *Human Genetics* Oct 109(4) (2001) 463.

[13] Faivre L., Vekemans M., Risk factors for heart defects in Down syndrome, *Teratology*, Mar, 59(3) (1999) 132.

[14] Fayyad U.M., Piatetsky-Shapiro G., Smyth P., Uthurusamy R. (red.), *Advances in Knowledge Discovery and Data Mining*, AAAI/MIT Press, Cambridge Mass. 1996.

[15] Fayyad U.M., Piatetsky-Shapiro G., Smyth P., Uthurusamy R., From data mining to knowledge discovery, in [14], pp.1-36.

[16] Fixler D.E., Threlkeld N., Prenatal exposures and congenital heart defects in Down syndrome infants, *Teratology*, 58 (1998) 6-12.

[17] Freeman S.B., Taft L.F., Dooley K.J et al, Population-based study of congenital heart defects in Down syndrome, *American Journal of Medical Genetics*, Nov 16;80(3) (1998) 213-7.

[18] Garcia A.M., Fletcher T., Benavides F.G., Orts E., Parental agricultural work and selected congenital malformations, *Am. J. Epidemiol.* 149 (1999) 64-74.

[19] Granzotti J.A., Paneto I.L. et al., Incidence of heart defects in Down syndrome, *Jornal de pediatria*, Jan-Feb;71(1) (1995) 28-30.

[20] Greco S., Matarazzo B., Słowiński R., Dealing with missing data in rough set analysis of multi-attribute and multi-criteria decision problems. in: [75],

[21] Greco S., Matarazzo B., Słowiński R., Handling missing values in rough set analysis of multi-attribute and multi-criteria decision problems. in: [75],

[22] Greco S., Matarazzo B., Słowiński R., Rough Sets Theory for Multicriteria Decision Analysis, *European Journal of Operational Research* 129(1) (2001) 1-47.

[23] Hastie T., Tibshirani R., Friedman J., The Elements of Statistical Learning; Data Mining, Inference, and Prediction, Springer -Verlag, New York 2001.

[24] James W.H., The male excess in Down syndrome, *Journal of Medical Genetics*, Sep; 33 (9) (1996) 806.

[25] Jansen R.P., The effect of female age on the likelihood of a live birth from one in-vitro fertilisation treatment, *Medical Journal of Australia*, May 17; 178 (6) (2003) 258-261.

[26] Kącki E., Kulikowski J.L., Nowakowski A., Waniewski E., *Systemy komputerowe i teleinformatyczne w służbie zdrowia*, Akademicka Oficyna Wydawnicza EXIT, Warszawa 2002.

[27] Kala R., *Statystyka dla przyrodników*, Wydawnictwo Akademii Rolniczej im. A. Cieszkowskiego w Poznaniu, Poznań 2002.

[28] Khoury M.J., Erickson J.D., Implications for the interpretation of increasing rates of cardiovascular malformations in surveillance systems, *American Journal of Epidemiology* Dec 15;136(12) (1992) 1457-64.

[29] Klosgen W., Żytkow J.M., Handbook of Data Mining and Knowledge Discovery, Oxford Press 2002.

[30] Kohavi R., Quinlan J.R., Decision Tree Discovery, in [29], pp. 267-276.

[31] Krawiec K., Stefanowski J., *Uczenie maszynowe i sieci neuronowe*, Wydawnictwo Politechniki Poznańskiej, Poznań 2003.

[32] Kubat M., Bratko I., Michalski R. S., Review of machine learning methods, in [39], pp. 3-70.

[33] Lapin L.L., *Probability and Statistics for Modern Engineering*, PWS Engineering, Boston, Massachusetts 1983.

[34] Leonard S., Bower C. et al., Survival of infants born with Down syndrome: 1980-96, *Paediatric and perinatal epidemiology*, Apr; 14 (2) (2000) 163-171.

[35] Leshin L., Trisomy 21: The Story of Down Syndrome, www.ds-health.com.

[36] Lin T.Y., Wildberg A. (eds), Soft Computing: Rough Sets, Fuzzy Logic, Neural Networks, Uncertainty Management, Knowledge Discovery, Simulation Councils Inc., San Diego, 1995.

[37] Loffredo C.A., Hirata J. et al., Atrioventricular septal defects : possible etiologic differences between complete and partial defects, *Teratology*, Feb; 63(2) (2001) 87-93.

[38] McCullagh P., Nelder J.A., *Generalized Linear Models*, Chapman and Hall Ltd, London 1983.

[39] Michalski R.S., Bratko I., Kubat M. (eds), *Machine learning and data mining*, John Wiley & Sons, 1998.

[40] Michie D. (eds), *Expert systems in the micro electronic age*, Edinburgh University Press 1979.

[41] Miletic T., Aberle N. et al, Perinatal outcome of pregnancies in women aged 40 and over, *Collegium antropologicum*, Jun; 26 (1) (2002) 251-8.

[42] Mokhtar M.M., Abdel-Fattah M., Major birth defects among infants with Down syndrome in Alexandria, Egypt (1995-2000): trends and risk factors, *East-Mediterr-Health-Journal*, May;7(3) (2001) 441-451.

[43] Pal S.K., Skowron A. (eds), *Rough fuzzy hybridization. A new trend in decision-making*, Springer-Verlag, Hongkong 1999.

[44] Pawlak Z., Rough sets, International Journal of Information & Computer Sciences 11 (1982) 341-356.

[45] Pawlak Z., *Rough Sets: Theoretical Aspects of Reasoning about Data*, Kluwer Academic Publishers, Dordrecht 1991.

[46] Pawlak Z., Słowiński K., Stefanowski J., Teoria zbiorów przybliżonych w analizie danych medycznych, in [26], pp. 253-269.

[47] Pawlak Z., Słowiński R., Rough set approach to multi-attribute decision analysis. *European Journal of Operational Research* 72 (1994) 443-459.

[48] Phillips P.C., The Exact Distribution of the Wald Statistic, *Econometrica*, 54 (4) (1986) 881-895.

pp. 146-157.

pp. 295-316.

[49] Quinlan J. R., Discovering rules by induction from large collection of examples, in [40].

[50] Quinlan J.R., C4.5: Programs for Machine Learning, San Francisco, Morgan Kaufmann, 1993.

[51] Romero-Maldonado S et al, Effects of risk on the child of an older mother: a case control study, *Ginecologia y obstetricia de Mexico*, Jun; 70 (2002) 295-302.

[52] Rose 2 -Rough Set Data Explorer - User's Guide, Laboratory of Intelligent Decision Support Systems, Institute of Computing Science, Poznan University of Technology, http://www-idss.cs.put.poznan.pl/rose.

[53] Schneider D., The Heart and Children with Down Syndrome, www.pirchei.co.il/specl_ed/down/archives/heart.htm

[54] Shashi V., Berry M.N., Covitz W., A combination of physical examination and ECG detects the majority of hemodynamically significant heart defects in neonates with Down syndrome, *American Journal of Medical Genetics*, Mar 15;108 (3) (2002) 205-8.

[55] Słowiński K., Stefanowski J., Medical Information Systems -problems with analysis and way of solution, in [43], pp. 301-315.

[56] Słowiński K., Stefanowski J., Siwiński D., Application of Rule Induction and Rough Sets to Verification of Magnetic Resonance Diagnosis, *Fundamenta Informaticae* 53 (2002) 345-363.

[57] Słowiński R., Od sztucznej inteligencji do sztucznego życia, czyli o aktywnej funkcji informatyki, *Pro Dialog* 16 (2003) 51-76.

[58] Słowiński R., Rough Set Approach to Decision Analysis, *AI Expert Magazine* 10 (3) (1995) 18-25.

[59] Słowiński R., Stefanowski J., Rough set reasoning about uncertain data, *Fundamenta Informaticae*, 27 (2-3) (1996) 229-244.

[60] STATISTICA (data analysis software system), version 6, StatSoft, Inc. (2001), www.statsoft.com.

[61] Stefanowski J., *Algorytmy indukcji reguł decyzyjnych w odkrywaniu wiedzy,* Rozprawa habilitacyjna, Politechnika Poznańska, Instytut Informatyki, Poznań 2001.

[62] Stefanowski J., Urbaniak S., Using case-based learning in decision support systems.

[63] Texas Heart Institute, Congenital Heart Disease, www.tmc.edu/thi/congenit.html

[64] The Site for Health and Bioscience Teachers and Learners, www.accessexcellence.org

[65] Torfs C.P., Christianson R.E., Maternal Risk Factors and major associated defects in infants with Down syndrome, *Epidemiology*, May,10(3) (1999) 264-270.

[66] Tubman T.R., Shields M.D. et al, Congenital heart disease in Down`s syndrome: two year prospective early screening study, *British Medical Journal*, 302 (1991) 1425-1427.

[67] Venugopalan P., Agarwal A.K., Spectrum of congenital heart defects associated with Down Syndrome in high consanguineous Omani population, *Indian pediatrics*, May; 40 (5) (2003) 398-403.

[68] Webster's New Medical Dictionary, Wiley Publishing, Inc, 2003.

[69] Weiss S.M., Kulikowski C.A., *Computer Systems That Learn*, Morgan Kaufmann, San Mateo 1991.

[70] Wilk Sz., *Flexible, Knowledge-Based Decision Support Systems in Mobile Environment*, Rozprawa doktorska, Politechnika Poznańska, Instytut Informatyki, Poznań 2003.

[71] Yang Q., Rasmussen S.A., Friedman J.M., Mortality associated with Down`s syndrome in the USA from 1983 to 1997: a population-based study, *Lancet*, Mar 23; 359 (9311) (2002) 1019-1025.

[72] Zanakis S.H., Doukidis G., Zopounidis C. (eds): *Decision Making: Recent Developments and Worldwide Applications,* Kluwer Academic Publishers, Dordrecht, Boston 2001.

[73] Zespół ds. Polskiego Rejestru Wad Wrodzonych, Web Site of the Polish Registry of Congenital Malformations, www.registry-cong-malf.com

[74] Zespół ds. Polskiego Rejestru Wad Wrodzonych, Wrodzone wady rozwojowe w Polsce w latach 1998-1999. Dane z Polskiego Rejestru Wrodzonych Wad Rozwojowych, Ośrodek Wydawnictw Naukowych, Poznań 2002.

[75] Zhong N., Skowron A., Ohsuga S. (eds): *New Directions in Rough Sets, Data Mining and Granular-Soft Computing*, LNAI 1711, Springer-Verlag, Berlin 1999.