SEARCHING FOR CREDIBLE RELATIONS IN MACHINE LEARNING

Vedrana Vidulin

Doctoral Dissertation
Jožef Stefan International Postgraduate School
Ljubljana, Slovenia, January 2012

Supervisor: prof. dr. Matjaž Gams, Jožef Stefan Institute, Ljubljana, Slovenia
Co-supervisor: prof. dr. Bogdan Filipič, Jožef Stefan Institute, Ljubljana, Slovenia

Evaluation Board:
prof. dr. Marko Bohanec, Chairman, Jožef Stefan Institute, Ljubljana, Slovenia
assoc. prof. dr. Dunja Mladenić, Member, Jožef Stefan Institute, Ljubljana, Slovenia
prof. dr. Jurij Tasič, Member, Faculty of Electrical Engineering, University of Ljubljana, Ljubljana, Slovenia

MEDNARODNA PODIPLOMSKA ŠOLA JOŽEFA STEFANA
JOŽEF STEFAN INTERNATIONAL POSTGRADUATE SCHOOL
Ljubljana, Slovenia

Vedrana Vidulin

SEARCHING FOR CREDIBLE RELATIONS IN MACHINE LEARNING

Doctoral Dissertation

ISKANJE VERODOSTOJNIH RELACIJ V STROJNEM UČENJU

Doktorska disertacija

Supervisor: prof. dr. Matjaž Gams

Co-supervisor: prof. dr. Bogdan Filipič

Ljubljana, Slovenia, January 2012

To my family: Dino, Antica and Marčelo

Contents

Abstract

Can a model constructed by machine learning or data mining programs be trusted? For example, it is known that a decision tree model can contain less-credible parts caused by pathologies in induction algorithms, noise and missing values in data, or simply because of the complexity of a domain. Such models typically contain relations that are statistically significant, but in reality meaningless. Meaningless relations are problematic since they undermine the user's trust in the data mining system and can also lead to wrong conclusions about the most important relations in the domain.

In this thesis we propose an interactive method for the construction of credible relations in complex domains, named Human-Machine Data Mining (HMDM). The basic idea of our approach is to construct a large number of models to extract the credible relations, i.e., relations that are meaningful and of high quality. The task is computationally very demanding, and for other than simple cases there is no possibility for humans to analyze a meaningful share of all the hypothesized models on their own. However, the introduced combination of human understanding and raw computer power enables a smart examination of the parts of the huge search space with most credible models. While data mining methods perform the search, humans examine and evaluate the results, make conclusions and redo the search in a way that seems to be the most promising based on the previous attempts. In this way, the humans guide the data mining to search the subspaces with the most credible models and finally the humans construct the overall conclusions from the various, most interesting solutions.

The HMDM defines a toolbox composed of semi-automated data mining procedures and a set of scenarios for the human to guide the analysis towards credible models. Furthermore, it defines a scheme for the extraction of credible relations from multiple models, which provides support to the human analyst in the process of constructing correct conclusions about the domain.

The proposed approach is demonstrated in two complex domains that show how the higher education and the research and development sectors are related to economic welfare. In addition, we showed in a domain of automatic web genre identification that HMDM can be successfully used for learning predictive models in another domain.

A user study justified the HMDM method by showing that the users are frequently not able to detect meaningless relations by observing a single model constructed by a machine learning algorithm. However, by observing interesting variations, i.e., candidate solutions suggested by the HMDM method, the participants realized the weaknesses of the default model and created better domain models.

Povzetek

Ali je mogoče zaupati modelu, zgrajenem z algoritmi strojnega učenja in rudarjenja podatkov? Znano je, da lahko model v obliki odločitvenega drevesa vsebuje slabe, tj. manj verodostojne dele, ki jih povzročajo patološko obnašanje indukcijskih algoritmov, šum in manjkajoče vrednosti v podatkih, lahko pa se pojavijo tudi zaradi kompleksnosti domene. Takšni modeli vsebujejo relacije, ki so statistično na videz pomembne, vendar v resnici vsebinsko nepomembne. Take relacije spodkopavajo zaupanje uporabnikov v sistem za rudarjenje podatkov in lahko privedejo do napačnih sklepov o najpomembnejših relacijah v domeni.

V disertaciji predlagamo interaktivno metodo za gradnjo verodostojnih relacij v kompleksnih domenah, ki jo poimenujemo Metoda rudarjenja podatkov človek-stroj (angl. Human-Machine Data Mining – HMDM). Osnovna ideja našega pristopa je, da zgradimo veliko število modelov, iz katerih pridobimo verodostojne relacije, ki so smiselne in visoke kakovosti. Naloga je računsko zelo zahtevna in za vse primere, razen preprostih, ljudje brez pomoči računalnika ne morejo analizirati ustreznega deleža vseh možnih modelov. Vendar pa predstavljena kombinacija človeškega razumevanja in surove moči računalnika omogoča pameten pregled najpomembnejših delov ogromnega preiskovalnega prostora. Medtem ko metode rudarjenja podatkov preiskujejo, uporabniki preverjajo in vrednotijo rezultate, sklepajo in usmerjajo iskanje na način, ki se zdi najobetavnejši. Na ta način uporabniki usmerjajo proces rudarjenja podatkov proti pomembnim delom preiskovalnega prostora in na koncu gradijo zaključne sklepe iz različnih najzanimivejših rešitev.

Metoda HMDM definira nabor programskih orodij, ki vsebuje polavtomatske postopke rudarjenja podatkov in niz scenarijev, ki pomagajo uporabnikom, da vodijo analizo v smeri verodostojnih modelov. Poleg tega metoda določa način pridobivanja verodostojnih relacij iz več modelov, s katerim zagotavlja podporo analitiku v procesu gradnje pravilnih sklepov o domeni.

Predlagano metodo smo demonstrirali na dveh kompleksnih domenah, ki pojasnjujeta, kako sta sektor visokega šolstva in sektor raziskav in razvoja povezana z gospodarsko blaginjo. Poleg tega smo v domeni avtomatske identifikacije spletnih žanrov pokazali, da je metodo HMDM možno uspešno uporabiti tudi za učenje napovednih modelov iz druge domene.

S pomočjo uporabniške študije smo pokazali prednosti metode HMDM, ker uporabniki pogosto ne uspejo zaznati nesmiselnih relacij z opazovanjem enega samega modela, zgrajenega z algoritmom strojnega učenja. Vendar pa so z opazovanjem zanimivih variacij, t.j. možnih rešitev, ki jih predlaga metoda HMDM, uporabniki spoznali slabosti privzetega modela in posledično ustvarili boljše modele.

Abbreviations

ACC	=	accuracy
AWGI	=	automatic web genre identification
BRPT	=	binary relevance problem transformation method
CC	=	correlation coefficient
CCPE	=	corrected class probability estimate
CPE	=	class probability estimate
CPX	=	complexity
DM	=	data mining
EVD	=	extreme value distribution
FN	=	false negatives
FP	=	false positives
HMDM	=	Human-Machine Data Mining
IDM	=	interactive data mining
LRS	=	log-likelihood ratio statistic
MCP	=	multiple comparison procedure
ML	=	machine learning
MNIL	=	minimum number of instances per leaf
MNR	=	minimal number of relations in a model
PP	=	percentage points
R&D	=	research and development
RAA	=	relative absolute accuracy
REP	=	reduced error pruning
TP	=	true positives

1 Introduction

1.1 Credible Relations in Machine Learning

This thesis considers a specific problem related to data mining (DM) and machine learning (ML). DM comprises the automatic and semi-automatic processes of discovering patterns in data (Witten and Frank, 2005), while ML algorithms induce patterns from the data (Mitchell, 1997). The thesis further specializes in interactive data mining (IDM) (Fails and Olsen, 2003; Zhao and Yao, 2008; Zhao, 2008) as a subfield of DM with the emphasis on improved man-machine interaction supported by the DM and ML methods.

The thesis introduces a new method HMDM (Human-Machine Data Mining) for the extraction of credible relations from multiple models constructed with the DM and ML methods from given data. *Relations*, as addressed in this thesis, are patterns that connect a set of attributes/independent variables that describe the properties of a concept underlying the data, and a class/target attribute/dependent variable that represents the concept. For example, suppose that a concept is the economic welfare of a country. One of the tasks is to understand what differentiates rich from poor countries. If the attributes are the level of participation in education and the level of investment in the research and development (R&D) sector, the HMDM can lead to relations: the rich countries invest a lot in R&D sector, while the poor countries exhibit low levels of participation in education.

Definitions of credibility, meaning and quality are following:

Definition 1.1.1: *Meaning* is defined as a subjective criterion attributed by the human based on the common sense, an informal knowledge about the domain, observed frequency, strength and stability of the relation.

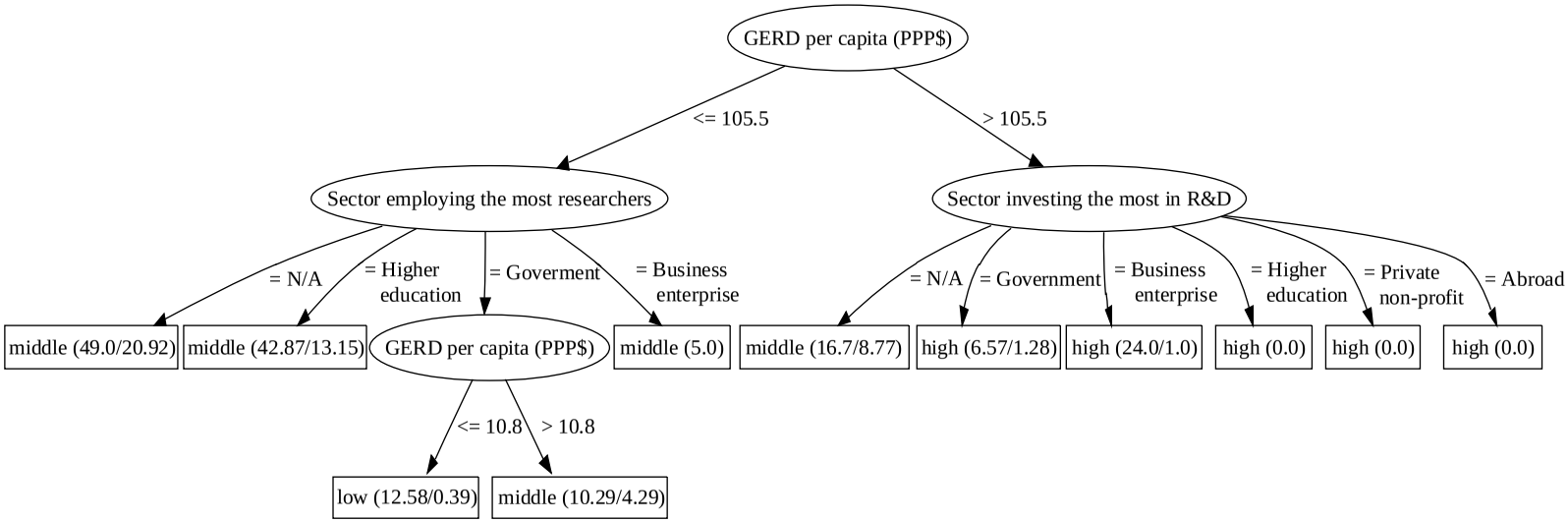
Definition 1.1.2: *Quality* represents an objective criterion that indicates a support of the selected quality measures.

Definition 1.1.3: A *credible* relation is of great meaning, i.e., meaningful, and of high quality. A relation is *less-credible* if it is either meaningless or of low quality or both.

Here, a model consists of relations, and a credible model consists of credible relations.

The HMDM method searches the search space of all relations in a domain by examining as many of them as possible using two criteria: meaning and quality, constituting credibility.

Relations and consequently models can be expressed in different languages, depending upon the knowledge representation that the DM and ML methods implement. Models induced by *black-box methods* are usually of high quality, but do not clarify which relations describe the concept. In contrast, models induced by *transparent-box methods*, such as classification rules and decision trees, transparently explain the induced concepts (Langley, 1996). However, transparent-box models can contain less-credible relations due to pathologies in the induction algorithms (Jensen and Cohen, 2000), missing values and the noise in data, and because of the complexity of the domain. Such models contain relations, which are statistically significant, but in reality meaningless.



ACC 64.67%; CCPE 0.4113; Kappa 0.4047; CPX 11

Figure 1.1: The decision tree constructed from the R&D attributes.

An example of less-credible model is the decision tree in Figure 1.1 that is constructed by default parameters and the minimum number of instances per leaf (MNIL) equal to 5 in Weka (Witten and Frank, 2005). The tree is constructed from the data set composed of 37 attributes describing R&D sector of a country, 167 examples representing countries and the class that differentiates countries according to their economic welfare into low, middle and high. Nodes in the tree represent attributes and leaves represent the class. Connections between the attributes and the class form the relations. At each leaf the first number in brackets represents the number of examples that reach that leaf. The second number represents the number of the examples that take the class value other than the one represented by the leaf. Quantities are expressed in decimals to account for weights of the examples with missing values. The tree contains three relations. The first states that countries with better welfare invest a lot in R&D – the attribute "GERD per capita (PPP$)" (GERD stands for Gross Domestic Expenditure on R&D and PPP$ for purchasing power parity in American dollars) that represents the level of investment in R&D appears twice in the tree and both times the "higher than" side of the subtree (> 10.8 and > 105.5) leads to leaves representing better welfare than the "less than" side. The expected role of the second relation including "Sector employing the most researchers" is to make clear distinction between "low" and "middle" countries. However, for "middle" countries any sector could be the main employer, which makes the relation meaningless. The third relation including "Sector investing the most in R&D" is also meaningless since the only "middle" leaf represents the countries for which the sector is unknown ("N/A" value). By examining the three relations, one can establish that the first relation is meaningful because it is a common sense relation that makes clear distinctions between the countries with different levels of welfare. However, the credibility of the relation would be increased by having an additional evidence that the tree constructed only from the "GERD per capita (PPP$)" attribute is of higher quality than the presented tree. Similarly, the other two attributes that form the meaningless relations in the presented tree, may form meaningful relations when combined with other attributes from the data set. The additional evidence is not available in a single constructed tree. For this purpose, the HMDM method, the main contribution of this thesis, constructs multiple models (decision trees in the presented example) in a specific manner in order to examine the contained relations for credibility and finally, to extract credible relations from the multiple constructed models.

To eliminate less-credible relations from a single model, different automatic approaches have been suggested. For example, post-pruning (Quinlan, 1993) is a common approach, which often eliminates too specific relations, but does not eliminate all less-credible relations, especially in models constructed for complex domains. Post-pruning is already applied in Figure 1.1; however, less-credible relations remained within the tree.

We propose an interactive method and a toolbox composed of semi-automated DM procedures for humans in order to extract credible relations from multiple models. Credible relations provide support to the human analyst in the process of constructing conclusions about the domain and form the credible models that improve the models constructed by automatic DM and ML methods.

1.2 Hypothesis and Purpose

The hypothesis of the thesis is that a combination of human understanding and raw computer power will enable a smart examination of the parts of the huge search space with most credible models. In this way, the user will construct models and, with the help

of the relation-extraction scheme that we propose, extract credible relations from multiple models.

The purpose of the dissertation is to enable humans to analyze complex domains, design and cross-check various relations and to extract credible relations. The domain is defined by data on which DM and ML tools can be applied. On top of this, a new IDM algorithm will make it possible to achieve the purpose.

The main goals of the dissertation are:

- Development of the HMDM method for the extraction of credible relations from complex domains.
- Establishment of the relation-extraction scheme.
- Testing and verifying a statistical measure corrected class probability estimate (CCPE) on decision trees.

Other goals are:

- The application of the HMDM method in three complex domains.
- Evaluation of the HMDM usability through a user study.
- Analysis of problems leading to less-credible relations in models.
- Survey of state-of-the-art IDM methods.

Note that the CCPE measure is also applicable to transparent-box models other than decision trees. However, in this thesis we only present the experiments performed on decision trees.

1.3 Scientific Contributions

We propose a new method, named HMDM, which extracts credible relations from multiple models constructed with DM and ML methods on given data. The method and analyses related to this thesis were published in journals and conference proceedings: (Vidulin and Gams, 2011; Vidulin et al., 2009, 2007b; Vidulin and Gams, 2006a, 2010, 2009; Vidulin, 2009; Vidulin and Gams, 2008a,b; Rehm et al., 2008; Vidulin and Gams, 2007; Luštrek et al., 2007; Vidulin et al., 2007d,a,c; Vidulin and Gams, 2006b; Vidulin et al., 2006). The complete bibliography is presented in Appendix E.

The main contributions of this thesis are the following:

- A new method Human-Machine Data Mining (HMDM) was developed for extracting credible relations from data, based on interactive and iterative process exploiting advantages of human and artificial intelligence (Chapter 4).
- The CCPE statistical measure, originally conceived for classification rules, was extended for decision trees (Section 4.4).
- Interactive explanations of DM results were designed, conceived to facilitate the extraction of credible relations (Chapter 5).

Additional contributions of the thesis:

- A computer program was developed to support the HMDM method.
- For two real-life domains, how the higher education sector and R&D sector influence the economic welfare of a country, we extracted credible relations with the new method, confirming some well-known relations and providing some new ones (Chapter 5).
- For the real-life domain of automatic web genre identification (AWGI), we constructed credible models with the new method, which provide an insight into the role of content words in recognizing web genres (Chapter 6).

1.4 Overview of the Thesis Structure

Chapter 2 presents a brief description of the ML task and introduces the terminology necessary to discuss this process. Models constructed by ML algorithms can contain less-credible relations caused by the pathological behaviour of the algorithms, especially when constructed for complex domains. Many of these relations appear as a result of an optimistic estimation of the model's quality. Chapter 2 illustrates less-credible relations through examples and presents a statistical explanation of the causes.

IDM methods designed to interact with the user in order to improve the models constructed by the DM and ML methods, mostly deal with two issues. First, how to explain the results of DM and ML to the user, and second, how to collect the corrective feedback. Chapter 3 presents the related methods, which address the two problems.

The main contribution of this thesis – the HMDM method – is described in Chapter 4. The algorithm that formalizes the steps of the method combines human understanding and raw computer power in order to extract credible relations that are meaningful and of high-quality at the same time. The main idea is to construct a large number of models in a specific way. The human examines these models, makes a conclusion, and directs the search towards the parts of a search space with credible models. At the same time, the human extracts the credible relations from multiple constructed models using relation-extraction scheme, which differentiates two types of relations (combination and redundancy) and divides them into three levels of credibility.

In addition, Chapter 4 describes several quality measures used to assess the quality of relations and models. In Section 4.4 we present the extension of the CCPE statistical measure.

The HMDM method is designed for domain-analysis tasks, where the goal is to construct the correct conclusions about the domain. In Chapter 5 we demonstratehow the HMDM is applied for designing and cross-checking relations with the goal to extract credible relations. The method is applied in two complex domains showing how the higher education and R&D sectors influence economic welfare of a country.

In Chapter 6 we present another application of the HMDM, which constructs predictive models for the domain of AWGI. This time, the credible models, composed only of credible relations, are constructed with the interactive HMDM algorithm and compared with the models constructed with an automatic model-induction algorithm. The results show that the credible models are superior in terms of both meaning and quality.

Chapter 7 presents the user study designed to test the usability of the HMDM method. The study showed that the participants found the approach beneficial in several ways.

The thesis concludes with Chapter 8, where we present discussion, our conclusions and ideas for future work.

2 Problems Leading to Less-Credible Relations in Models

This chapter first presents a succinct description of the basic concepts of ML, and introduces the terminology necessary to discuss the model-induction process. Next, the emphasis shifts to the possible causes of less-credible relations in models. Less-credible relations frequently occur due to the pathological behaviour of ML algorithms, especially when applied to complex domains. We present two examples of less-credible relations and a statistical interpretation of the pathologies leading to such relations.

2.1 Machine Learning

It is generally considered that the goal of a ML algorithm is to find a model that generalizes a set of observations. In this process, three components should be defined: a representation of the observations in computer-understandable form, a language for describing the model and an algorithm that constructs the model.

In ML terminology, observations are named *instances* or *examples*, and are represented in the form of a table named *data set*. A simplified example of the data set is presented in Table 2.1, where each row is an instance and each column an *attribute*. Attributes can have two functions. *Unlabelled attributes* describe the properties of the instances (attributes A_1 and A_2 in our example). In statistics, these attributes are called independent variables. In ML, they are typically referred to as attributes. *Labelled attributes* define target values of the instances (attribute C). In statistics, these attributes are named dependent variables. In ML, they are typically referred to as classes when the target values are discrete and as target attributes when the target values are continuous. In the case of *supervised learning*, the labelled attributes are defined and the goal is to learn a function $\mathbf{Y} = f(\mathbf{X})$, which predicts the labelled attributes $\mathbf{Y}$ using the unlabelled attributes $\mathbf{X}$. In the case of *unsupervised learning*, the labelled attributes are not defined and the goal is to find natural groupings within the data that are based on a similarity between the instances. In this thesis we deal with the supervised learning methods.

Table 2.1: An example of a data set that represents the concept of logical AND.

A_1	A_2	C
0	0	0
0	1	0
1	0	0
1	1	1

Models or *hypotheses* can be described in different languages, which are dependent upon the knowledge representation a ML algorithm implements. Three types of knowledge representation are the most common: *threshold concepts*, *competitive concepts* and *logical conjunctions* (Langley, 1996). Threshold concepts use weights to indicate the importance

of unlabelled attributes in describing each of the classes. A new instance belongs to a class when a sum of weighted attribute values for the instance exceeds a certain threshold. Typical examples of threshold concepts methods are neural networks and Naïve Bayesian classifiers. Competitive concepts are represented with sets of instances, one for each class. A new instance belongs to the class that contains the most similar instances in terms of attribute values. Typical examples are instance-based ML algorithms such as nearest neighbour. Logical conjunctions describe each class with one or several conjunctions of conditions that an instance should satisfy to belong to a certain class. Classification rules and decision trees are the typical representatives.

Transparent-box algorithms often use a logical conjunction type of representation. For example, the data set in Table 2.1, which represents the logical AND concept, can be described with the rule: IF $A_1 = 1$ AND $A_2 = 1$ THEN $C = 1$ ELSE $C = 0$.

Hypothesis space is the space of all the possible hypotheses that can be expressed by means of the selected representation. For example, in the case of a decision tree induction problem, the hypothesis space represents all decision tree models that can be constructed from given data. The ML algorithm searches the hypothesis space with the goal to find the hypothesis that best generalizes the data. ML algorithms differ in the choice of a searching strategy and an evaluation function used to assess the quality of the examined hypotheses (Mitchell, 1997). We will focus on the algorithms that construct transparent-box models and discuss possible causes of less-credible relations in the models related to the choice of the evaluation function. But first we will illustrate the less-credible relations in models with two examples.

2.2 Examples of Less-Credible Relations in Models

2.2.1 The Great Horse-Manure Crisis of 1894

Everyday predictions are commonly based on false assumptions, which lead to an over- or under-estimation of the real consequences. An example of the overestimated prediction is: if the trend X continues, the disaster is going to happen. Such a prediction was made in 1930s, when it was predicted that most Western countries were about to enter a terminal decline. Ten years later, the baby boom happened. Predictions after the baby boom were concerned with overpopulation. In recent decades, with fertilities of around 1.5, publications report about a European demographic "black hole". In contrast, some predictions underestimated the real effect. For example, in the 19th century it was predicted that by 1950 every town in America would have a telephone. Davies (2004) argued that the main cause of such false predictions lies in an assumption that the events will continue to occur in the same way as they did in the past.

A typical example of overestimation, often cited in economic literature, is the problem of the horse-manure crisis that emerged at the end of the 19th century (Davies, 2004). All of the transport of that time was horse-powered. For example, in 1900 in London there were 11,000 cabs and several thousand buses, where each bus required 12 horses. On average 50,000 horses were used daily just for the transport of people. In addition, horses were also used for the transportation of goods. Similar statistics could be obtained for other big cities of that time (McShane and Tarr, 2007). Considering that a horse, on average, produces between 7 and 16 kg of manure per day, the population of 100,000 horses produced approximately 1,134 tons of manure per day (Burrows and Wallace, 1999). Based on these numbers and the increasing demand for horses in the growing cities, the journalist of the Times of London in 1894 estimated that in 50 years the streets

of London would be buried under 2.7 meters of manure. However, the prediction was wrong, since the crisis vanished when motor vehicles replaced horses.

2.2.2 The Role of Females in Science

When the prediction models are constructed with ML algorithms, the overestimation usually results in models that contain seemingly high-quality relations, which are in reality meaningless. An example of such a model is presented in Figure 2.1. The decision tree model was constructed as a part of our analysis of which segments of the R&D sector have the highest impact on the economic welfare of a country. The model is constructed with the J48 algorithm (Witten and Frank, 2005), an implementation of C4.5 (Quinlan, 1993). We suspected that the increased level of investment in R&D ("GERD per capita") and the number of applications for patents represent more valid causes of better economic welfare since several publications indicated this to be so. One might also suspect that the level of participation of females is more a consequence of the increased economic welfare than a possible cause. There are, however, no direct clues in the model that can support or contradict our default assumptions. Since more important attributes are mostly positioned higher in the tree, one can only assume that "GERD per capita" is important, and it also most consistently appeared in the trees that were constructed with different parameters.

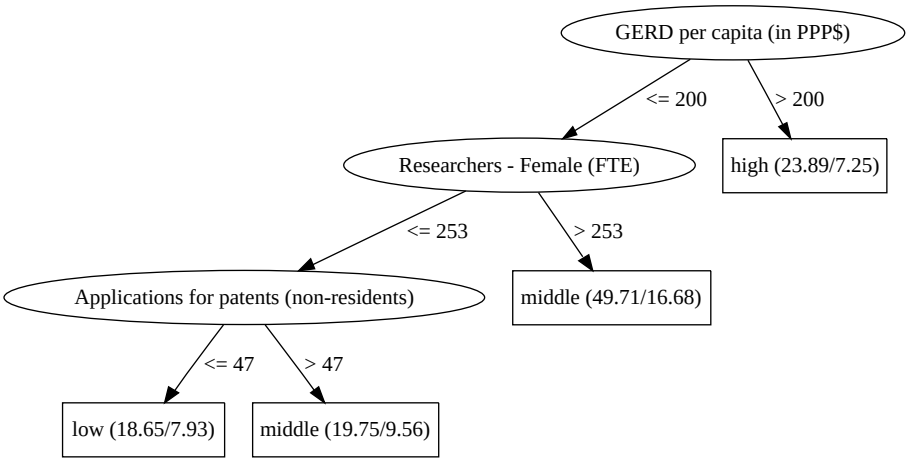


Figure 2.1: The decision tree constructed from the attributes representing the R&D sector, with the J48 algorithm from Weka.

The method proposed in this thesis offers formal support to the examination of suspicious relations in models for credibility and the extraction of the credible relations from the multiple models. With the help of our method, we were able to show that the increase in the level of investment in R&D and the number of patents indeed represent important prerequisites for improving economic welfare, while the participation of females is of less importance.

2.2.3 Discussion

The two examples show the difficulties behind the inductive reasoning that can lead to less-credible relations in models. Important attributes may not even be included in the analysis. For example, the journalist of the Times did not account for the increase in the price of horse-drawn transport, which resulted in incentives for people to find alternatives. This type of problem is difficult to address. There are rarely simple causes, and it is easy

to overlook an attribute from the current point of view. Leaving aside this pitfall, if we suppose that all important attributes are present in the data, the question remains as to whether they are correctly denoted by the ML algorithm as such. As can be seen from the second example, even less important attributes can appear in the model. In the next section we will present a statistical explanation of this phenomenon.

2.3 Optimism of Evaluation Functions

In a model construction process, the choices of models and model's components (e.g., a subtree, an attribute-value pair) rely on the scores computed by an evaluation function. The scores are computed on a *train set* – a random sample drawn from a population, which makes the score a statistic, and the decisions based on the score statistical processes of the parameter estimation and hypothesis testing. Jensen and Cohen (2000) argued that less-credible relations in models are caused by wrong assumptions behind the statistical procedures, which are explicitly or implicitly embedded in the induction algorithms. The algorithms typically compare multiple components based on scores and select the component with the maximum score to be included in the model. They termed the procedure a *multiple comparison procedure* (MCP). The problem arises because the algorithms do not adjust for the specific statistical properties of the MCP.

The MCP problem is best illustrated with an example. Suppose that we want to construct a model from a given sample of 15 instances. We wish to obtain a high-quality model, and consequently define that an acceptable model is one that would correctly classify 12 or more of the 15 instances. We expect that such model would contain relations that generalize over the whole population and not only over the specific sample. To assess the validity of our strategy, we compute the probability of observing a random model that would guess correct classes for 12 or more instances, using equations presented in (Dekking et al., 2005). For this purpose, we first define R_i, where $i = \{1, 2, \ldots, 15\}$:

$$R_i = \begin{cases} 1 & \text{if the the } i\text{th instance is correctly classified} \\ 0 & \text{if the the } i\text{th instance is incorrectly classified.} \end{cases} \tag{2.1}$$

Then, we define a random variable X that denotes the number of correctly classified instances:

$$X = R_1 + R_2 + \ldots + R_{15}, \tag{2.2}$$

where X attains values $0, 1, \ldots, 15$.

In general, the number of ways in which the k out of n instances can be correctly classified is obtained by computing k-permutations of n:

$$C_{n,k} = \binom{n}{k} = \frac{n!}{k!(n-k)!}. \tag{2.3}$$

Considering that X has a binomial distribution with parameters $n = 15$ and $p = 1/2$ (the probability of a correct classification for a single instance is 50%), the probability that $X = k$ is computed as:

$$P(X = k) = \binom{15}{k} \left(\frac{1}{2}\right)^k \left(\frac{1}{2}\right)^{15-k} = \binom{15}{k} \left(\frac{1}{2}\right)^{15}. \tag{2.4}$$

Finally, the cumulative probability of observing a random model that would guess correct classes for 12 or more instances is equal to:

$$P(X \geq 12) = \binom{15}{12}\left(\frac{1}{2}\right)^{15} + \ldots + \binom{15}{15}\left(\frac{1}{2}\right)^{15} = 0.0176. \quad (2.5)$$

Since the probability is rather small (0.0176), we conclude that we can believe the relations within the model. After all, the chances to make an erroneous decision by accepting the model are no more than 0.0176.

Now, suppose that we constructed 10 different models from the sample (e.g., by changing the algorithm's parameters). Formally, we observed a sequence of 10 independent random events $X_1, X_2, \ldots, X_m$, where $m = 10$. Then, suppose that we select the model with the highest score and again apply the same validity test. This time the test is different, because the goal is to compute the probability that the best observed model would guess correct classes for 12 or more instances. For this purpose, we define a random variable $X_{max} = max\{X_1, X_2, \ldots, X_{10}\}$ and compute the probability using the formula:

$$P(X_{max} \geq k) = 1 - (1 - q)^m, \quad (2.6)$$

where, $q = P(X_i \geq k)$.

Applied to our example:

$$P(X_{max} \geq 12) = 1 - (1 - 0.0176)^{10} = 0.1627. \quad (2.7)$$

The resulting probability shows that in the new setup, the probability of observing a random model with such a high score is almost 10 times higher than in the case of a single model. If we were to examine 100 models, the situation would get even worse: $1-(1-0.0176)^{100}$ or 0.8306. By not accounting for the specific properties of the MCP, we can easily select a completely random model by assuming that it is 80% accurate (12/15 = 0.8), while in reality the performance of the model on the population would be no better than 50%.

A typical error with ML algorithms is that they treat the maximum score as an unbiased estimator of the model's quality. However, the score is unbiased as long as m, the number of models, is equal to one. For m greater than one, the maximum score overestimates the real value. The result is an optimistic estimation of the model's quality. In statistics, optimistic or conservative means that the real value of statistic (measured on a population) cannot be any better than the current estimate. For practical purposes of the model induction, this means that by examining more models we can more easily obtain a random model of high quality, in the case that we do not adjust for the MCP.

2.3.1 The Statistics of Optimistic Estimators

Lets state that there are m models, and for each a score x is computed on a sample S, using an evaluation function $x_i = f(model_i, S)$. The score x_i is statistical in nature, since different samples can produce different scores for the same model. In other words, x_i is a specific value of a *random variable* X_i. For a given f and a model, the values of x_i for all the possible samples of size $|S|$ drawn from a given population define the *sampling distribution* of X_i.

A maximum score x_{max} is computed from all of the m models:
$x_{max} = max\{x_1, x_2, \ldots, x_m\}$. x_{max} is a specific value of a random variable X_{max}. In

contrast to X_i, the sampling distribution of X_{max} depends on the number of examined models m.

Ideally, we would like that x_{max} is a good estimate of the model's population score ψ_*. In statistical terms, we would like that X_{max} is an unbiased estimator of ψ_*[1]. In general, an estimator X of a population parameter ψ is considered as unbiased if $E(X) = \psi$, where $E(X)$ denotes an expected value of X. Jensen and Cohen (2000) constructed a proof that X_{max} is a positively biased estimator, based on a comparison of $E(X_{max})$ with ψ_*. The proof is composed of two parts. First, they established that $E(X_i) < E(X_{max})$. Then, they used this relationship to show that X_{max} is a biased estimator of ψ_*.

Theorem. *For discrete random variables $X_1, X_2, \ldots, X_m$, where all x_i are scores and $x_{max} = max\{x_1, x_2, \ldots, x_m\}$,*

$$E(X_i) \leq E(X_{max}). \tag{2.8}$$

Proof: The expected value of the discrete random variable X is defined as the sum, over all possible values x, of the value x multiplied by its probability $p(x)$:

$$E(X) = \sum_x xp(x). \tag{2.9}$$

For scores, each possible value x is derived from one or more samples S. Each sample produces only a single value x, although many samples may produce the same value x. Because of this many-to-one mapping from the samples S to the values x, the expected value of a discrete random variable can equivalently be defined over all possible samples S:

$$E(X) = \sum_S x(S)p(S). \tag{2.10}$$

where $x(S)$ is the value of x for a given sample S.

Given that the function max selects among the values $x_1, x_2, \ldots, x_m$, for any score x_i, $x_i \leq max\{x_1, x_2, \ldots, x_m\}$, where $1 \leq i \leq m$. More succinctly, $x_i \leq x_{max}$. For a given population, x_i and x_{max} are summed across the same samples, and these samples have identical probability distributions. Therefore,

$$E(X_i) \leq E(X_{max}). \tag{2.11}$$

If for one or more samples, $x_i < x_{max}$, then

$$E(X_i) < E(X_{max}). \tag{2.12}$$

Theorem. *Given a sample S and a corresponding ψ_*, the population score of the model with the maximum sample score,*

$$\psi_* \leq E(X_{max}). \tag{2.13}$$

for $m > 1$. That is, X_{max} is a biased estimator of the population score ψ_*.

Proof: If every X_i is an unbiased estimator of the population score ψ_i, then

$$\psi_i = E(X_i). \tag{2.14}$$

[1] ψ_* should not be confused with ψ_{max}, since $\psi_{max} = max\{\psi_1, \psi_2, \ldots, \psi_m\}$.

As previously proven, $E(X_i) \leq E(X_{max})$. Thus, for all ψ_i

$$\psi_i \leq E(X_{max}). \tag{2.15}$$

If, for one or more samples, $x_i < x_{max}$, then

$$\psi_i < E(X_{max}). \tag{2.16}$$

That is, X_{max} is a positively biased estimator of any ψ_i, including the population score ψ_* of the item with the maximum sample score, so

$$\psi_* < E(X_{max}). \tag{2.17}$$

In other words, X_{max} is a biased estimator of ψ_*.

3 Related Work

A comprehensive overview of IDM is presented in (Zhao, 2008). In this thesis, however, we focus on two IDM issues: explanations of the DM results to the user and the inclusion of corrective feedback into the DM process.

3.1 Explanations

Explanation is "the act or process of explaining", as well as "something that explains" (http://www.merriam-webster.com). In the context of DM, the second sense is used. In this light, we can define the explanation as a knowledge representation used to facilitate an interpretation of the DM results. In the literature, the explanations are generally divided into two groups depending upon whether the goal is to interpret a model or a model's decision to classify an instance into a specific manner.

3.1.1 Instance-Based and Model-Based Explanations

An explanation of model's decision or instance-based explanation presents which attributes and to what extent influenced model's decision to classify an instance into a specific manner. In the case of the transparent-box models, such as decision trees, the explanation is constructed simply by highlighting the path in the tree that resulted in a specific classification (Quinlan, 1993). In contrast, the black-box models use tabular or graphical representation of the decision. For example, Kononenko (1993) presented the decision of a Naïve Bayesian classifier in tabular form by quantifying the influence that each attribute had in the decision process. Štrumbelj et al. (2009) used bar charts to show which attribute values contributed to the decision of an arbitrary classifier to classify an instance into a specific manner.

Explanations of model or model-based explanations again differ for transparent-box and black-box models. An explanation of a transparent-box model is typically constructed by simplifying the model. For example, Bohanec and Bratko (1994) simplified decision tree by pruning. In contrast, an explanation of a black-box model is constructed either by translating the model into the transparent-box model or by visualizing model's relations. An example of translation is presented in (Towell and Shavlik, 1993), where a neural network is translated into a rule set. Similarly, Craven (1996) translated the same type of model into a decision tree. An example of visualization is presented in (Becker et al., 2001), where pie and bar charts are used to visualize the Naïve Bayesian classifier in order to show the ranges of the attribute values that characterize a given class. Similarly, Možina et al. (2004) used nomograms, a graphical representation of numerical relations, as an explanation of the Naïve Bayesian classifier.

Several of the approaches simultaneously implement both types of explanations. Poulin et al. (2006) presented ExplainD, a framework that exploits the bar charts to explain both the decision and the model for classifiers that use additive evidence. Lacave and Diez (2002) presented a review of different types of explanations constructed for Bayesian

networks, categorizing them into the explanation of the evidence, the explanation of the model and the explanation of the decision.

3.1.2 Static and Interactive Explanations

The literature further divides the explanations into static and interactive. A common feature of the static explanations described up to this point is an assumption that the underlying model is a correct domain description. There are two problems connected with this assumption. First, if the model contains less-credible relations and the user overlooks them, he/she can make wrong conclusions about the most important relations in the domain. Second, if the user identifies the less-credible relations, but does not have a mean to remove them, this can undermine user's trust in the DM system (Stumpf et al., 2009). The relatively new approach of interactive explanations aims to address the presented problems by providing tools for the user to improve the model. For this purpose, Stumpf et al. (2009) conducted a user study and showed that a feasible interactive explanation should enable the user to change the importance/weight of the attributes that are highlighted by the model. An explanation is based on an instance being classified. Kulesza et al. (2009) implemented the idea of Stumpf et al. (2009). Like with static explanations of the Naïve Bayesian classifier's decisions, the impact of the most important attributes is presented with the help of bar charts. However, in contrast, the user can expand/reduce a bar to increase/decrease the contribution of an attribute in recognizing a certain class.

The main goal of the presented interactive explanations is to improve the predictive performance of the underlying model. The explanations are instance-based and the user does not need to review and reason about the complete model, which can be complex and even contain less-credible relations that slightly improve the predictive performance, but do not contribute to the meaning of the model. Our goal is different. Since our task is a domain analysis, we propose an interactive explanation that supports the user in the process of constructing correct conclusions about the domain. The conclusions are made based on credible relations, credibility of which is established by examining the relations' role within multiple models as explained in the following chapter. The process is supported by the model-based explanation, which in contrast to the related approaches, encompasses multiple models arranged in a manner to facilitate the extraction of credible relations. The user interacts with the explanation in order to extract the credible relations and to store credible models, composed only of meaningful and high-quality relations.

3.2 Inclusion of Corrective Feedback into the DM Process

Corrective feedback is a correction of a model provided by the user, which is typically collected using an interactive explanation of DM results. Two types of corrections are the most common: example-based and knowledge-based. In the case of former, the user corrects the model either by providing new training examples, or by providing the correct class for an incorrectly classified example. In the case of later, the user makes a direct correction on model's components, e.g., by changing the weights of attributes that participated in the classification of a specific instance (Stumpf et al., 2009).

Fails and Olsen (2003) proposed an approach where the user iteratively improves the classifier by generating new examples and retraining the system. Žnidaršič and Bohanec (2007) presented a model revision approach, where the model constructed by an expert is

refined based on a set of new examples provided by the user. Xiao et al. (2002) described a metasynthetic approach where the user observes classifications made by the system and provides the corrections in the case of an error. The corrected examples are then used to retrain a meta-classifier that combines the output of several sub-classifiers. MacKay (1992), Cohn et al. (1996), Tong and Koller (2002), and Melville et al. (2005) all presented an active learning approach where the system iteratively asks the user to provide a label for the most informative example until a satisfactory prediction performance is not reached. In terms of computational learning theory, in the presented approaches the role of the user is that of an oracle who provides the learning system with the examples (Kearns and Vazirani, 1994).

Another set of approaches treats the corrective feedback as a special case of introducing domain knowledge into the ML process. In contrast to the example-based approaches, the corrections directly address the model's structure, ensuring a more rapid improvement. Culotta et al. (2006) looked at the ML process as an optimization process, where corrections are introduced as constraints. Similar approaches were introduced by Shilman et al. (2006) to improve handwriting recognition and by Huang and Mitchell (2006) to refine the results of clustering. Filipič et al. (1999) proposed a combined ML and genetic algorithm approach, where a genetic algorithm is used to improve a decision tree model by optimizing its numerical parameters. Stumpf et al. (2009) presented a user co-training approach, where two classifiers are co-trained – one representing the ML system's view of the data and the other representing the user's view of the data. Ware et al. (2001) described a tool for the manual construction of a decision tree in accordance with the user preferences, supported by a data-visualization technique. Zhao and Yao (2005) presented a tool for the construction of a granule tree (similar to the decision tree), where the user selects a granule/attribute according to his/her preferences. The user's decision is supported by several measures of the granule's quality.

The approaches presented so far are all focused on the interactive improvement of a single model. An alternative is to generate multiple models and to select one or several that are the best. Nguyen et al. (2000) presented a model-selection system called CABRO where the user constructs several decision trees by changing the parameters of a tree-induction algorithm and then selects the best model by inspecting the constructed trees. Osei-Bryson (2004) presented a multi-criteria, decision-analysis approach for the extraction of a set of the best decision trees, where the user states the preferences as the weights of five performance measures and as the constraints imposed on the ranges of the measures' values.

We propose a constraint-based, multi-criteria, model-selection approach, where the user constrains the DM process by defining the parameters and attribute subsets of interest. The model-construction process within the initial DM (see Chapter 4) is a combination of the ideas presented in (Nguyen et al., 2000) and (Osei-Bryson, 2004).

4 Human-Machine Data Mining

The basic idea of our approach is to construct a large number of models to extract the credible relations. The task is computationally very demanding because from n binary attributes it is possible to construct 2^{2^n} decision theories. The space of all the potential hypotheses for 100 binary attributes and a single binary class is therefore $2^{2^{100}}$. This number is far larger than the number of all the atoms in our universe, which according to Wikipedia is around 10^{80}, i.e., 2^{266}. Therefore, for other than simple cases humans cannot analyze any meaningful share of all the hypothesized models without the help of automatic or semi-automatic methods for finding promising hypotheses.

We propose the method that combines human understanding and raw computer power in order to enable a smart examination of the parts of the huge search space with most credible models. When DM methods perform a search, the humans can examine and evaluate the results, make conclusions and redo the search in a way that seems to be the most promising based on previous attempts. In this way, the humans guide the DM to search the subspaces with the most credible models and finally the humans, in their mind, construct the overall conclusions from the various most interesting solutions.

4.1 Basic Ideas

Our approach is based on two assumptions: first, that from the enormous number of all the hypothesized models only a couple of them best represent the domain in terms of meaning and quality (in analogy to the Occam's razor); and second, that the search mechanism will discover the credible models. Therefore, optimal or suboptimal models will be constructed by the DM and recognized by humans as such. Indeed, this is our experience in recent years in most of the real-life domains describing social and economic relations.

We use two basic heuristics: first, we examine the whole set of various parameters (typically, algorithm parameters and attribute subsets) to get a clue about where the credible models might be; and second, as soon as a candidate model occurs, several heuristics are applied for cross-checking the credibility of the observed and similar models. The models confirmed as credible are stored, and a new search begins until no major new models are found for a while.

The proposed method for the extraction of credible relations from multiple models can be illustrated with a simple example. Let us state that there is a data set D containing 10 attributes $A_1, \ldots, A_{10}$, a class/target attribute C and two credible relations: A_1 AND A_3 and A_7 OR A_8. The two relations represent patterns that connect the attributes from the relations with the class/target attribute C.

With our method the former relation is established by constructing three models from: A_1 and C, A_3 and C, the combination of the two attributes A_1 and A_3, and C. Note that later C will not be explicitly presented in relations.

When the first and second models are of poor quality, while the third model is of considerably better quality, at the same time containing meaningful relations that include

both attributes, the relation type A_1 AND A_3 can be established. The details of this relation can also be further established, but at this point the explanation might blur the simple example. The relation is referred to as a **combination** and is denoted as $A_1\&A_3$.

The A_7 OR A_8 relation is also established by comparing three models, but the procedure is different. First, a model M is constructed from D. Second, a model M' is constructed from D', obtained by removing A_7 from D. At this point, two conditions indicate the A_7 OR A_8 relation: a) A_8, which represents the same semantic category as A_7, takes the role of A_7 within the structure of M'; and b) M' is of a quality similar to M. If at least the first condition is satisfied, the following step is to remove A_8 from D' (resulting in D'') and to construct a model M'' from D''. If M'' is of a quality lower than M' and the relation is meaningful, the relation type A_7 OR A_8 is established. The relation is referred to as a **redundancy** and is denoted as $A_7||A_8$.

In addition, we used the heuristic function proposed in (Jakulin, 2005) to further assist the user in the relation-extraction process. Let us state that S is a subset of attributes from D, e.g., $\{A_1, A_3\}$. The interaction between the attributes in S is measured by the function presented in Eq. 4.1.

$$I(S) = -\sum_{T \subseteq S} (-1)^{|S \setminus T|} \left(-\sum_{v \in T} P(v) \log_2 P(v) \right). \tag{4.1}$$

The $P(v)$ represents a (joint) probability distribution for the values of the attribute(s) within the subset $T \subseteq S$. A positive interaction indicates a synergy between the attributes, indicating a combination. A negative interaction indicates overlapping between the attributes, indicating redundancy.

Furthermore, the proposed method divides the relations into **three levels of credibility**. A relation belongs to the first level when: a) the attributes from the relation form semantically similar structures within multiple models; b) the presence/absence of the attributes from the relation causes an increase/decrease in the model's quality over multiple models (see Section 4.4 for quality measures). With respect to the second criterion, a relation $A_1\&A_3$ is a candidate for the first level of credibility if, e.g., a model constructed from A_1, A_2, A_3 and C would be of a higher quality than a model constructed from A_2 and C, and a model constructed from A_1, A_3, A_5 and C would be of a higher quality than a model constructed from A_5 and C. In other words, the quality measures should be monotone with respect to the first level relations.

The second level contains less stable relations. For example, the relation $A_1\&A_3$ would belong to the second level if, e.g., the model constructed from A_1, A_2, A_3 and C would be of a quality higher than the model constructed from A_2 and C, while the model constructed from A_1, A_3, A_5 and C would be of a quality lower than the model constructed from A_5 and C. The second-level relations frequently describe subgroups within the data.

The third level contains all the other relations that are of lower quality and consequently less interesting to the user, but still often constructed by DM.

In summary, the method we propose extracts specific relations of two types, namely combination and redundancy, and of three levels of credibility. The type and credibility levels are categories of the relation-extraction scheme named accordingly *type-credibility scheme*.

4.2 The HMDM Algorithm

The credible relations are extracted from the data with the help of the procedure presented in Figure 4.1, and are arranged in the type-credibility scheme described in the previous section. Within the procedure, the user begins the analysis by supplying the data of interest and by constructing an initial model. Based on the initial model, the user decides to examine the credibility of one or several of the model's relations. This is achieved through an interactive search guided with the help of the REMOVE_ATTRIBUTES and ADD_ATTRIBUTES procedures, as well as with the expand credibility indicator tool (Section 4.3). The results of the search are evaluated with several quality measures (Section 4.4), as well as with the heuristic function presented in Eq. 4.1. Based on the observed evidence, the user makes conclusions about the credibility of the examined relations and decides about further searching steps. When there are more interesting models and relations in the search space near the initial model, the user continues with the interactive search. In contrast, when all interesting models and relations are examined, the user stores the credible models and relations and integrates conclusions with the results of previous analyses. Finally, the user decides whether to continue the analysis by selecting another initial model or to conclude the analysis.

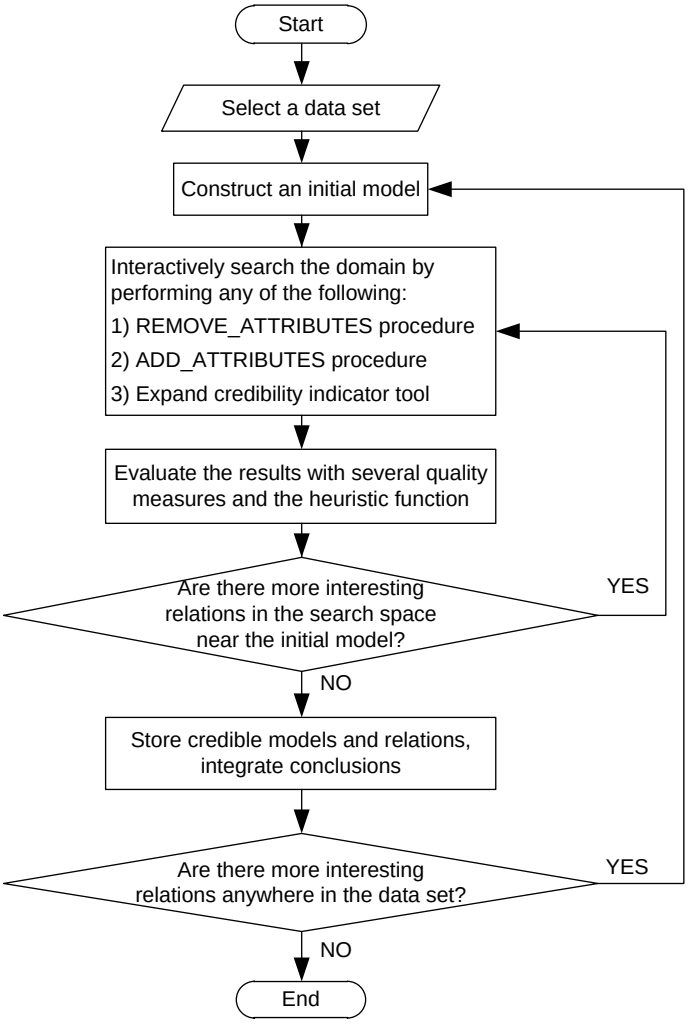


Figure 4.1: The flowchart presenting a top-level description of the HMDM algorithm.

The flowchart in Figure 4.1 presents a top-level description of the method we call Human-Machine Data Mining (HMDM). The HMDM algorithm that formalizes the steps of the method is presented in Algorithm 4.1. The basic step is a typical DM search that consists of choosing an interesting subset of attributes, and performing the DM with various parameters. Unlike ending with the best model, as is common in a DM search, HMDM relies on human decisions to choose one or a couple of interesting candidate models, re-analyzing them by changing the parameters and attributes, and repeating the loop until the relations within the model(s) are confirmed or rejected. Too many variations of the search options cause a combinatorial explosion; however, this search is guided by a human goal to verify the already-found interesting relations. As seen from the examples in Chapter 5, several variations can be quickly discarded as unpromising and the interesting ones can get human attention, demanding a reasonable time consumption.

Algorithm 4.1: The HMDM algorithm.

HMDM

1 Select a data set
  REPEAT
2 Modify attribute set
3 Select DM method
4 Select parameters and their ranges, define constraints
5 Perform INITIAL_DM(data set, DM method, parameters and ranges, constraints), creating a list of models LM
6 FOR each interesting model M from LM, reexamine M:
    REPEAT
     Perform any of the following: {
      REMOVE_ATTRIBUTES(data set, DM method, M, parameters and ranges, constraints)
      ADD_ATTRIBUTES(data set, DM method, M, attribute subset, $\varnothing$, parameters and ranges, constraints)
      Expand credibility indicator }
     Evaluate the results with several quality measures and with the heuristic function
    UNTIL no more interesting relations are found in the search space near the initial model
7   Store credible models and relations, integrate conclusions
  END FOR
  UNTIL no more new interesting relations are found anywhere in the data set

The steps of the HMDM algorithm are as follows:

Step 1: The user selects a data set.

Step 2: Attribute set modifications are optional and include: a) the construction of new attributes, e.g., by applying sum, min, max or ratio functions on two or more numerical attributes; b) attribute subset selection based on expert knowledge or executed by means of an arbitrary automatic method.

Step 3: The user selects one or several of the DM methods. We used the term "DM method" to emphasize that the user can choose a method that encompasses DM steps other than model construction. However, the method is suitable only if it satisfies two criteria: first, that the method is capable of producing a model in a human-understandable

form (the decision and regression trees in our case); and second, that the method is designed for supervised learning problems. For any chosen DM method the human-computer session is performed until the user has not explored all the interesting models.

Step 4: By defining parameters and constraints the user defines preferences about the hypothesis subspace searched by the DM method.

Step 5: This step represents an exploratory data-analysis phase (Mardia et al., 1979). The INITIAL_DM procedure constructs all the possible models with the selected DM method and the defined parameters that satisfy the constraints. The models are sorted using the predefined criteria. This step resembles the construction part in the random forest algorithm (Breiman, 2001), with the difference being that the user defines the way the new models are constructed. The user observes and selects one or several interesting models.

Step 6: Each model marked as interesting is a starting point in the search for credible relations. The search is performed by modifying the models with attributes, either interactively with the expand credibility indicator tool, or by computer routines from the HMDM toolbox. The two types of modifications are the remove and the add attributes, which are described in Section 4.3. These can be carried out automatically by the procedures REMOVE_ATTRIBUTES and ADD_ATTRIBUTES, or interactively with the expand credibility indicator tool. The modified models are evaluated with several quality measures.

Step 7: The models and relations marked as credible by the user are stored.

The role of the human in the presented steps is following:

Step 1: The user supplies the data set, which describes a domain of interest, expressed in computer-understandable form.

Step 2: Attribute set modification step is typically omitted when the data is analyzed for the first time. During the preliminary analysis, the user may hypothesise that certain attributes should be better expressed in different form. Then, in the second cycle, he/she constructs new attributes to test the hypothesis. In contrast, when the user establishes that certain attributes are not interesting for further analysis, he/she can eliminate those attributes.

Step 3,4: The user defines the preferences about the hypothesis space he/she wants to examine. For example, decision trees with two or more relations.

Step 5: The user examines the models constructed by the INITIAL_DM procedure in decreasing order of sorting criteria. The models that contain interesting relations are selected for further analysis.

Step 6: For each model selected for further analysis, the user establishes the credibility of model's relations and relations within the models found in the vicinity of the initial model. To accomplish this task, the user first removes attributes starting with the attributes from the initial model and continuing with the attributes that appear in the models constructed from the reduced attribute sets. By applying the type-credibility scheme on the constructed models, the user establishes and extracts the credible relations, as well as credible models composed only of those relations that proved to be credible. In the next step, the user adds to an empty set different combinations of attributes that emerged during attribute removal and constructs models from the selected attribute sets. By applying type-credibility scheme on the newly constructed models, the user obtains an additional evidence, which is used to re-establish the type and credibility level of the extracted relations. Furthermore, the user can establish new relations by examining arbitrary combinations of attributes that seem interesting from the user's perspective.

During the attribute removal and addition, the user applies the heuristic function in Eq. 4.1 to attributes from a selected relation to obtain an additional evidence that confirms or rejects the established relation's type.

Step 7: In this step, the user integrates the conclusions based on the credible relations and models, with the conclusions made from the previous analyses of the same data.

In summary, the HMDM algorithm contains the following components:

1. the collection of the ML and DM algorithms, such as Weka (Witten and Frank, 2005) or Orange (Demšar et al., 2004), used for constructing models from the data (in our case decision and regression trees),
2. procedures: INITIAL_DM, REMOVE_ATTRIBUTES and ADD_ATTRIBUTES,
3. tools: expand credibility indicator,
4. standard routines from the HMDM toolbox, such as cross-validation (Kohavi, 1995), attribute selection, etc.

The procedures and tools will be further detailed in the following section.

4.3 Modifications

To extract relations from multiple models and to establish their credibility, new models are constructed from attribute sets that are modified by removing and adding attributes. By observing the models, the user discovers interesting relations. Then, for each interesting relation, the user collects multiple evidence considering both meaning and quality in order to establish relation's type and level of credibility. In the following two subsections, we will present an interactive explanation that facilitates relation-extraction and evidence-collection processes. Furthermore, we will explain the procedures for collecting the evidence.

4.3.1 Remove Attributes

When an attribute or a set of attributes is removed and the model reconstructed on the basis of the reduced data is of lower quality, the relation that contains the attribute(s) gains in credibility. These attributes are credible. In contrast, when the removal of an attribute or a set of attributes results in an equal or higher-quality model, the relation that contains the attribute(s) losses in credibility (e.g., a random binary attribute can decrease the decision tree's accuracy (ACC) by 5 to 10% (John, 1997)). These attributes are less-credible.

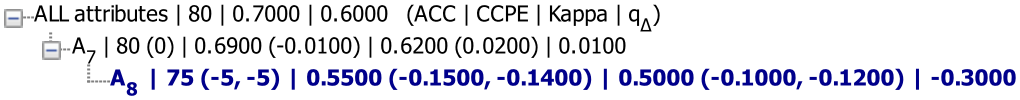


Figure 4.2: An example of the removed attributes graph.

The models constructed by removing attributes are arranged using an interactive explanation structure named the *removed attributes graph*. An example of the removed attributes graph is presented in Figure 4.2. The root of the graph represents the initial model, which is constructed from all attributes. The three numbers divided by the vertical bar represent quality measures obtained for the initial model. In this case, three measures are used: ACC, CCPE and Kappa. They are described in Section 4.4. The

presented measures can be substituted with any other set of measures, where higher numbers represent better models. The nodes in the graph are named *credibility indicators*, since they indicate the level of credibility. The credibility indicator at the first level of the graph represents the model constructed from the attributes that remained after the removal of the attribute A_7, as well as the credibility of the relation that contains the attribute A_7. The numbers that follow the vertical bar represent the quality of the model represented by the credibility indicator. The numbers in brackets that follow the quality measures represent the differences in quality in comparison to the initial model. These numbers serve for tracking changes in quality obtained by removing attributes. The last number represents the summary measure of quality q_Δ, which indicates the total change in quality caused by modifying the initial model (see Section 4.4 for details). The q_Δ measure presents a coarse-grained criterion for determining relation's credibility, while the differences in brackets represent the fine-grained criteria. For example, a fall in q_Δ indicates that the relation is promising and should be further examined, while the fall in all the quality measures further increases the relation's credibility. Finally, the credibility indicators at the deeper levels of the graph represent the removal of multiple attributes in combination: current attribute together with all of the superordinate attributes. Within the deeper levels, the second number in brackets represent the change in quality in comparison to the superordinate model. This number provides further evidence for establishing the relation's credibility. For example, the removal of the most credible attributes should reduce the quality of both the initial and superordinate models.

From Figure 4.2, it can be determined that the attribute A_7 is less-credible on its own, since q_Δ is positive and only one quality measure (CCPE) was reduced by removing the attribute. However, the graph further reveals that the relation composed of the two attributes A_7 and A_8 is credible – when two attributes are removed together, the quality considerably falls (q_Δ of -0.3 and the fall in all the quality measures). Therefore, the credibility indicator containing A_8 is written in blue, indicating the credible relation.

The next step is to establish the type and details of the relations that emerged as credible. The type is established by examining the models connected with the relation. For example, suppose that the attributes A_7 and A_8 are semantically similar and that the user observed that A_8 took the role of A_7 in the model constructed by removing A_7. Considering the fall in quality caused by the removal of both attributes, the redundancy relation $A_7||A_8$ can be established, as described in Section 4.1. At this point, the type is further re-examined with the heuristic function in Eq. 4.1. The interaction is computed between the attributes within the relation and the class/target attribute. For example, the negative interaction computed for the relation $A_7||A_8$ would increase the relation's credibility, and vice versa.

The details of the relation are established from the context in which the relations appear in several models. At this step the user attributes the meaning to the relation based on a common sense, an informal knowledge about the domain and observed frequency, strength and stability of the relation through multiple models. Based on the details, the user makes final judgement about the relation's credibility. For example, if the user knows from the literature that developed countries have more mobile phones per capita than developing countries, then the relation of opposite direction is not credible, even it may appear as credible from other evidence observed during the analysis.

The removed attributes graph can be constructed automatically with the help of the REMOVE_ATTRIBUTES procedure, interactively with the help of the expand credibility indicator tool and by combining the two approaches (we applied the combined approach, unless stated otherwise). The REMOVE_ATTRIBUTES procedure presented in Figure 4.3 constructs several credibility indicators, beginning with the attributes from the initial

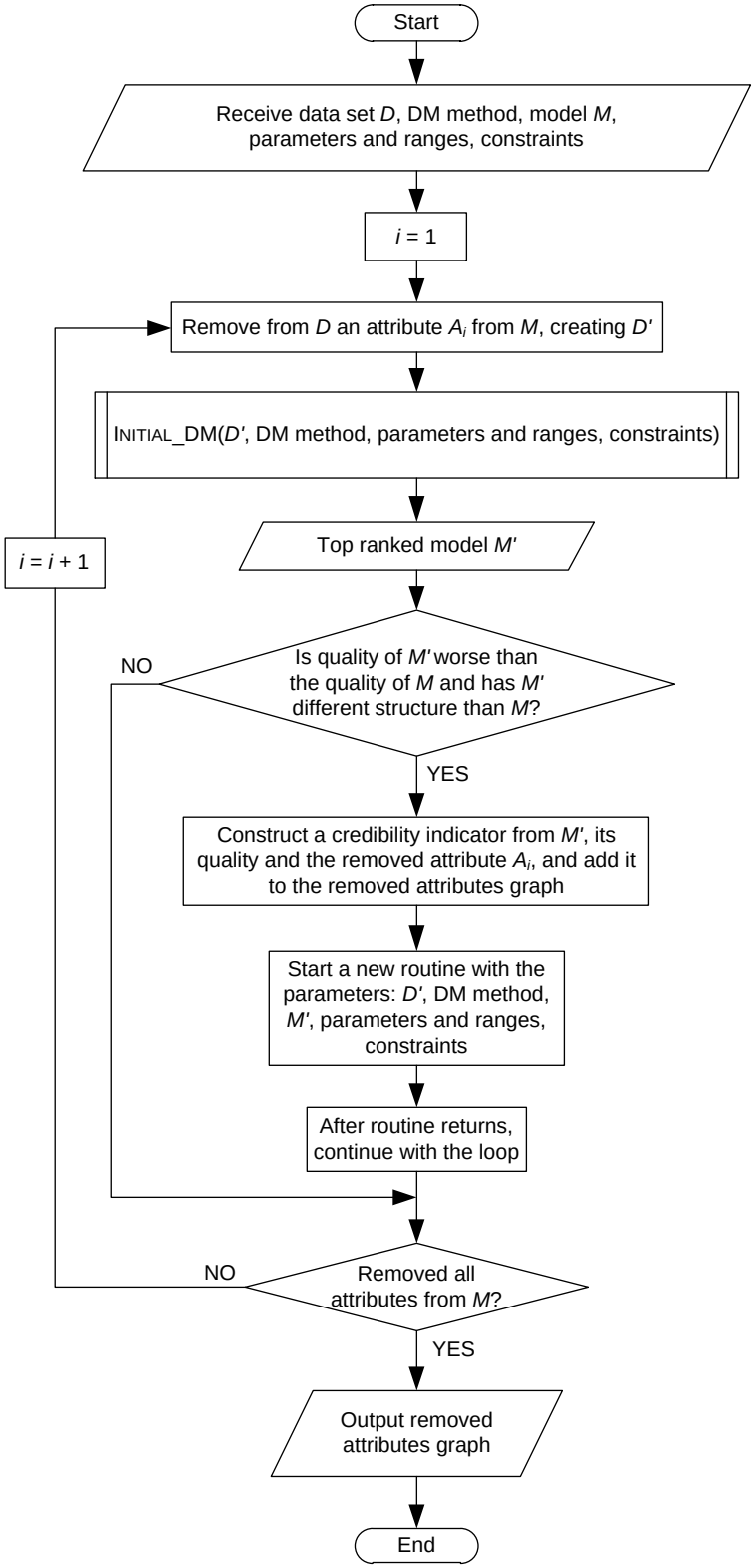


Figure 4.3: The flowchart of the REMOVE_ATTRIBUTES procedure.

model and removing them step by step. A new indicator is added to a current indicator if the quality of a newly constructed model M' decreases after eliminating the attribute and the structure of M' is different to the structure of the superordinate model M. The procedure resembles a wrapper attribute selection approach (Kohavi and John, 1997); however, our aim is to present the credibility of relations and models constructed from different attribute subsets to the user instead of finding a single best attribute subset.

Interactive construction of the graph is realized with the help of the expand credibility indicator tool. With this interactive tool, the user can refine the removed attributes graph, according to his/her preferences, by selecting an attribute or a set of attributes that are going to be removed. The credibility indicators are then automatically constructed.

4.3.2 Add Attributes

The basic idea behind adding attributes is to establish the credibility of a relation by detecting how the attributes within influence model's quality and structure in isolation from other attributes and relations. Frequently, when constructing a model, an attribute does not appear in the model, but may appear in one or more cross-validation folds, consequently influencing the quality of the model. In this manner, the user cannot establish the details of the relation, assess how meaningful the relation is and make the final judgement about the relation's credibility.

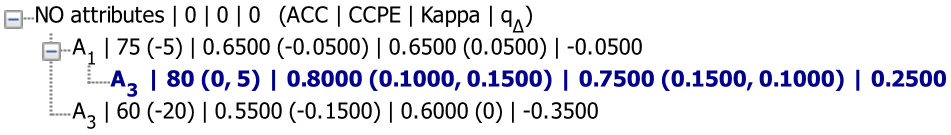


Figure 4.4: An example of the added attributes graph.

The relations' credibility is established using an *added attributes graph*, an example of which is presented in Figure 4.4. The added attributes graph resembles the removed attributes graph in terms of structure and function, with three differences. First, a root represents an empty attribute set. Second, the user searches for such attribute sets that improve the quality of the superordinate and the initial model. For example, when the two attributes A_1 and A_3 are separately added, the result are two models of quality lower than the initial model. However, the model constructed from the combination of the two attributes (blue indicator) is of considerably better quality, improving both the superordinate and the initial model. If the two attributes form meaningful relations within the improved model, a combination relation $A_1\&A_3$ (presented in Section 4.1) is established. Finally, the graph automatically identifies when the model does not contain all of the attributes from the selected attribute set. An indicator that represents such combination of attributes is strikethrough. Such combinations are useless, since the user cannot establish the type and details of the relation and, consequently, cannot determine the credibility of the relation.

Similar to removed attributes graph, the added attributes graph can be constructed either interactively with the help of the expand credibility indicator tool, or automatically. The ADD_ATTRIBUTES procedure presented in Figure 4.5 systematically constructs credibility indicators in reverse order compared to REMOVE_ATTRIBUTES, starting with an empty set and then gradually adding attributes. There are two differences between the two procedures. First, a new attribute is added to a credibility indicator when the quality increases after adding the attribute and the added attribute is present within the model's structure. Second, the user selects a set of attributes to be added: from the initial model, the removed attributes graph or simply by selecting an arbitrary set.

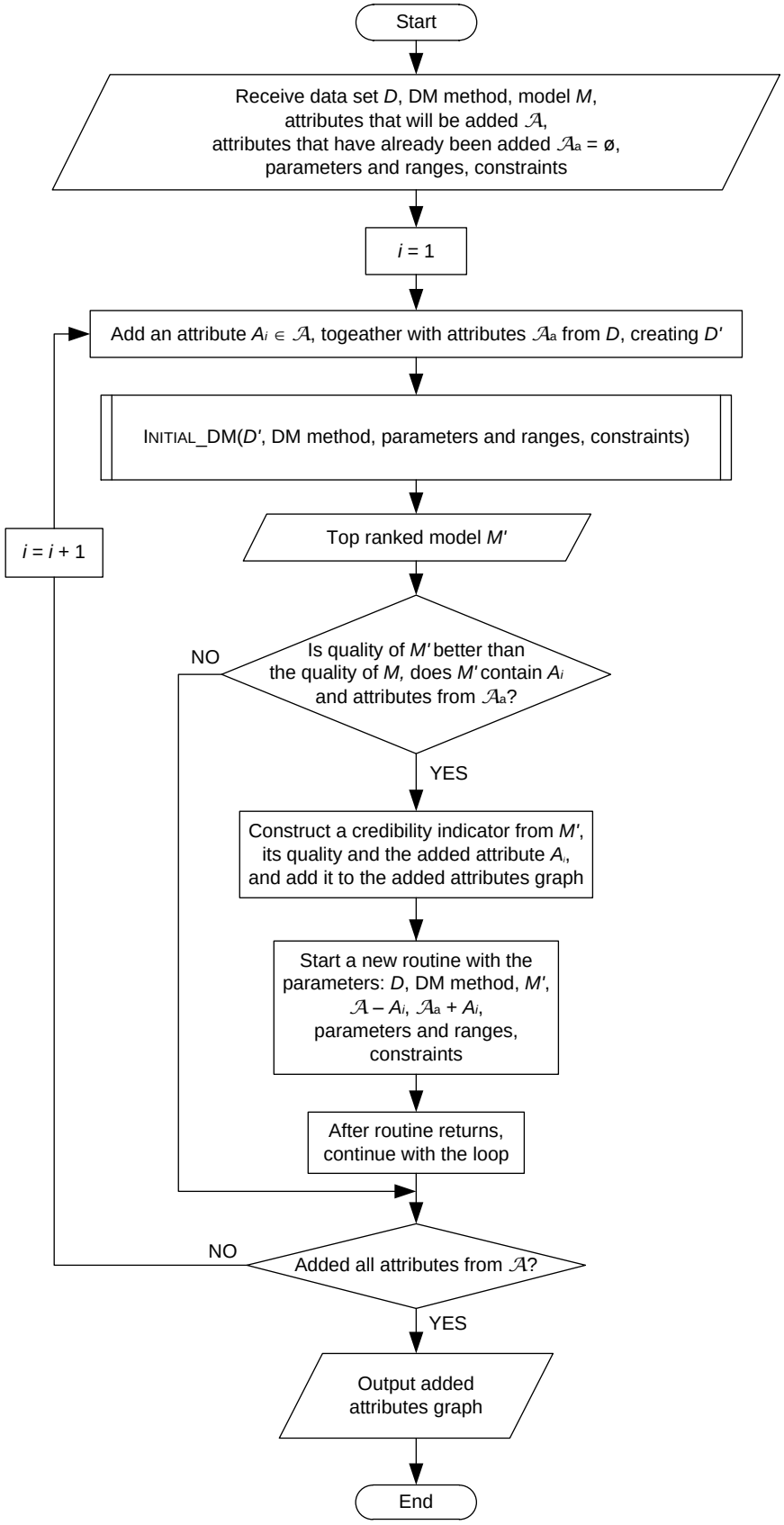


Figure 4.5: The flowchart of the ADD_ATTRIBUTES procedure.

4.4 Quality Measures

The quality of the classification models (in our case decision trees) is estimated with three measures. First, the accuracy (ACC) or success rate denotes the overall performance of a classifier, expressed as a percentage of correctly classified examples with respect to all the examples classified by the classifier. A higher value of ACC denotes a better classifier. Second, the corrected class probability estimate (CCPE) reflects the classifier's significance in comparison to all possible classifiers constructed on the same data. The values of the CCPE are distributed within the [0,1] interval, 0 indicating the worst, and 1 the best classifier that can be constructed for given data. Third, Cohen's Kappa indicates whether the agreement between the classifier's predictions and the actual class values exceeds the chance level. The values of Kappa are distributed within the [−1,1] interval, where 0 indicates a random classifier and 1 the best classifier. Values lower than 0.5 are undesirable.

The quality of the regression models (in our case regression trees) is estimated with two measures. First, similar to the ACC, the correlation coefficient (CC) denotes the overall performance of a regression model, expressed as a statistical correlation between the predicted and the actual target attribute's values. The CC is distributed within the [−1,1] interval, where "negative values should not occur for reasonable prediction models" (Witten and Frank, 2005), a CC of 0 denotes the worst, while a CC of 1 denotes the best model constructed from given data. Second, similar to Kappa, the relative absolute accuracy (RAA) indicates how much the model exceeds the chance level. The RAA is computed by subtracting the standard measure of the relative absolute error from 100, to obtain the measure where higher values denote a better model.

The quality measures are typically computed using a 10-fold cross-validation with a random seed equal to 1. All the measures, except the CCPE, which we adjusted to operate on decision trees, represent the standard in an evaluation of the classification and regression models (Witten and Frank, 2005); therefore, only the CCPE is explained in detail.

The procedure for computing the CCPE is originally designed for classification rules (Možina et al., 2006). The CCPE corrects the standard class probability estimate (CPE). CPE or relative frequency of the rule r (Eq. 4.2) is a proportion of examples correctly classified by the rule (s) with respect to all of the examples covered by the rule (n).

$$CPE(r) = \frac{s}{n}. \tag{4.2}$$

The CCPE is prone to assigning optimistically high values to random patterns in the data. This optimism is reduced by subtracting the proportion of those rules that are better than a constructed rule from the CPE of that rule.

To compute the CCPE for a decision tree, we observe the tree as a group of relations/rules. For each relation, the probability of the majority class in the leaf – the CPE – is computed first. The CPE is then corrected by means of the Fisher-Tippet extreme value distribution (EVD) (Fisher and Tippett, 1928), which represents all the possible trees constructed from the data. Finally, the CCPE for each relation is weighted by the proportion of examples covered by the relation and summed to obtain the CCPE for the tree. The described procedure employs two routines: one for computing the EVD and another for computing the tree's CCPE.

Algorithm 4.2: A procedure for computing the μ parameter of the EVD.

COMPUTE_EVD (data set D, the size of LRS sample)
Let max. decision tree depth $d = 1$
DO
DO
Permute values of class in $D \rightarrow D_P$
Learn a decision tree on D_P with LRS as evaluation measure and max. depth $= d$
Record LRS of the best relation
WHILE (predefined size of LRS sample is not reached)
Compute the μ parameter of the EVD for the depth d
$d = d + 1$
WHILE ($\mu(d) > \mu(d-1)$)
Return the list of μ parameters for different depths

The routine for computing the EVD is presented in Algorithm 4.2. The statistic sampled to compute the EVD is a log-likelihood ratio statistic (LRS), computed using the equation:

$$LRS = 2[s \log \frac{s}{e_s} + (n-s) \log \frac{n-s}{e_{n-s}} + s^c \log \frac{s^c}{e_{s^c}} + (n^c - s^c) \log \frac{n^c - s^c}{e_{n^c - s^c}}], \tag{4.3}$$

where s denotes the number of the majority class examples that reached the leaf, n is the number of all the examples that reached the leaf, s^c is the number of the majority class examples that did not reach the leaf, and n^c is the number of all examples that did not reach the leaf. e_s, e_{n-s}, e_{s^c} and $e_{n^c - s^c}$ are expected values, computed using the following equations:

$$e_s = n \frac{s + s^c}{n + n^c}, \tag{4.4}$$

$$e_{n-s} = n(1 - \frac{s + s^c}{n + n^c}), \tag{4.5}$$

$$e_{s^c} = n^c \frac{s + s^c}{n + n^c}, \tag{4.6}$$

$$e_{n^c - s^c} = n^c(1 - \frac{s + s^c}{n + n^c}). \tag{4.7}$$

The EVD has two parameters: μ representing the location and β representing the scale. For the LRS, β is always equal to 2. Therefore, the routine computes only the μ parameter, which is dependent upon the maximum depth of the tree. For each depth the μ is computed by constructing a predefined number of trees from given data (1000 in our case) under the assumption that there is no relation between the attributes and the class. For each tree the LRS of the best relation is sampled and μ is computed as a median of the LRS sample + 2ln ln2. The trees within the routine are constructed by the modified J48 algorithm from Weka, which we modified to: a) use the LRS instead of the entropy;

Algorithm 4.3: A procedure for computing the CCPE of a decision tree.

COMPUTE_CCPE (a decision tree)
FOR each relation in the decision tree
Compute s, n, s^c, n^c
Compute LRS
Compute area P under the EVD(μ(relation depth), β=2) with the LRS as a lower bound
Compute expected value of LRS ($\widetilde{LRS}$) by finding the lower bound for the area under $\chi^2(1)$ equal P
Compute the expected value of s ($\tilde{s}$) from the ($\widetilde{LRS}$) by a root finding algorithm
$CCPE = CCPE + \frac{\tilde{s}}{n} \times \frac{n}{n+n^c}$
END FOR
Return $CCPE$

b) complete the construction of the tree at the predefined maximum depth. Finally, the EVD is used to compute the tree's CCPE with the routine presented in Algorithm 4.3.

Within the removed and added attributes graphs, the effect of the modification is assessed through the summary measure of q_Δ, which indicates the total change in quality caused by modifying the initial model.

Let ACC_Δ be the difference in the ACC between a given modification and the initial classification model, $CCPE_\Delta$ be the difference in the CCPE and $Kappa_\Delta$ in the Kappa, then q_Δ is computed as:

$$q_\Delta = (\frac{ACC_\Delta}{100} + CCPE_\Delta + Kappa_\Delta). \tag{4.8}$$

To reduce all the measures to the same scale: a) ACC was divided by 100; and b) negative values of Kappa were reduced to 0. Consequently, the values of the q_Δ are distributed within the $[-3, 3]$ interval. A q_Δ of 0 indicates that there is no difference between the modified and the initial model, denoting a non-promising modification. The interpretation of positive and negative values for q_Δ differs based on whether the attributes are being removed or added. When removing attributes, the promising modification is the one with a negative q_Δ). In contrast, when adding attributes, the promising modification is the one with a positive q_Δ. Similarly, for regression models, q_Δ represents a linear combination of CC and RAA, where a negative CC is reduced to 0 and RAA is divided by 100:

$$q_{\Delta_{reg}} = (CC_\Delta + \frac{RAA_\Delta}{100}). \tag{4.9}$$

In summary, in this chapter we presented the HMDM method for the extraction of credible relations from multiple models constructed with the DM and ML methods from given data. The main advantage of the method lies in the interaction between the human that poses informal knowledge about the domain and the computer supported with the DM and ML methods. In this manner, the computer provides a formal evidence that indicates the relation's credibility, while the human makes final judgement based on meaning and the presented evidence.

In this chapter we applied the HMDM method on artificial examples. In the following chapter two applications of the HMDM method will be presented that show the applicability of the method in real-life domains. The presented analyses will further reveal the properties of the method.

5 Domain Analysis with the HMDM

The HMDM method is designed to facilitate the analysis of complex domains, such as macroeconomic (Vidulin and Gams, 2006a) and demographic (Gams and Krivec, 2008) domains. In this chapter, two applications of the HMDM method are presented, where the goal is to acquire credible knowledge about a domain of interest. The applicability of the HMDM is demonstrated for the two DM methods that construct decision and regression trees and two analyses showing the impact of the higher education (Section 5.1) and R&D sectors (Section 5.2) on economic welfare. The chapter concludes with an evaluation of the credible trees constructed during the analyses (Section 5.3).

5.1 The Impact of the Higher Education Sector on Economic Welfare

We collected the data representing higher education sector from two statistical databases provided by the UNESCO Institute for Statistics (http://www.uis.unesco.org) and USAID – Global Education Database (http://ged.eads.usaidallnet.gov). The economic welfare is represented by the "GNI per capita" attribute, calculated according to The World Bank Atlas method. GNI stands for the Gross National Income and represents the total value of goods and services produced within a country (Black et al., 2009). We collected the "GNI per capita" from The World Bank database (http://www.worldbank.org) in both numerical and discrete forms. The numerical form is expressed in US$, while the discrete form represents the official classification of the countries into income levels: low ($745) or less, middle ($746-9,205), and high ($9,206) or more. From the total of 167 countries, 50 belong to the low, 79 to the middle and 38 to the high income groups.

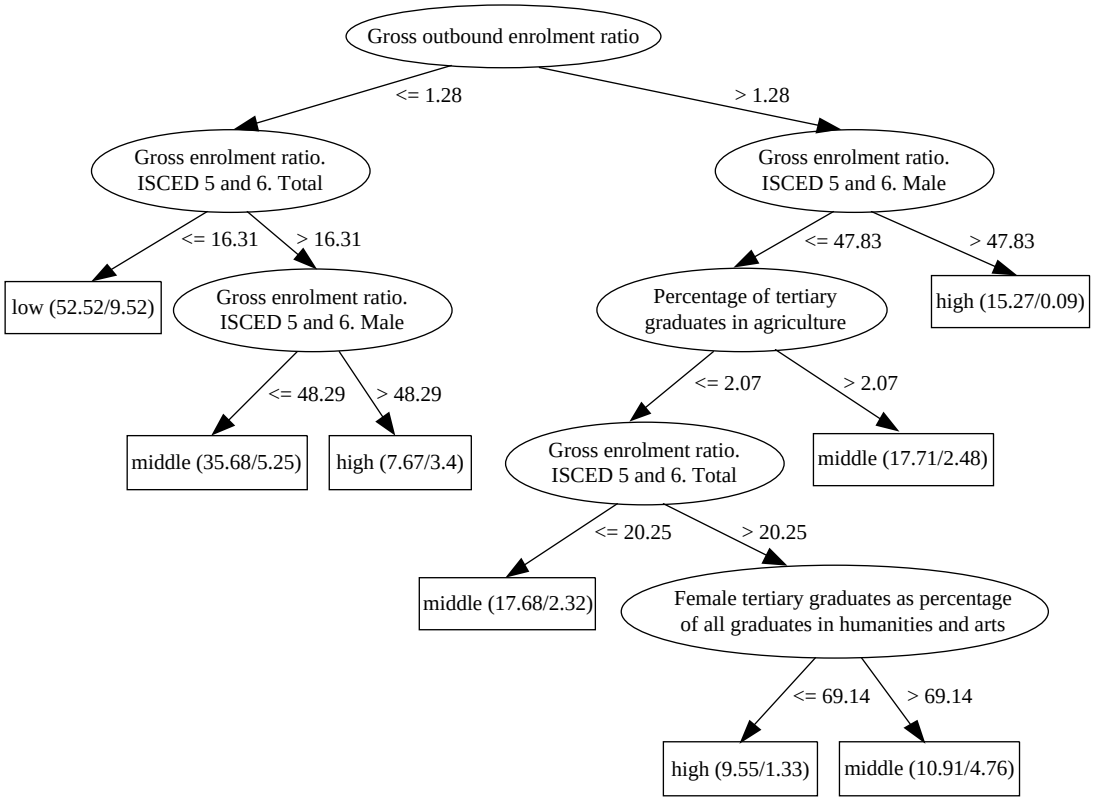
The data set is available at http://dis.ijs.si/Vedrana/economic-analysis.htm, while the description of attributes in provided in Appendix A. Note that abbreviated versions of the attributes that are used in the following subsections are presented in addition to the complete attribute names within Appendix A.

The following subsections present three HMDM analyses of the higher education data, performed by constructing: first, decision trees from the original attribute set (Subsection 5.1.1); second, decision trees from the modified attribute set (Subsection 5.1.2); and third, regression trees from the modified attribute set (Subsection 5.1.3). It is worth noting that only the most important findings are presented, although we considered a broad spectrum of evidence during the analyses.

5.1.1 Decision Trees Constructed from the Higher Education Data

The analysis is divided into steps, which exactly correspond to the steps of the HMDM method.

Step 1: The data set is composed of 60 attributes describing the higher education sector, discrete "GNI per capita" class, and 167 examples representing countries.
Step 3: Decision trees are constructed with the J48 algorithm from Weka, the implementation of C4.5.
Step 4: Two parameters of the J48 that control the model's complexity are selected: minimum number of instances per leaf (MNIL) with values ranged from 2 to 15 and reduced error pruning (REP) with on/off values. Two constraints are set: "minimal number of relations in a model" (MNR) is set to 2 and "remove duplicate models" is set to "on".
Step 5: The INITIAL_DM returned 14 trees, from which we selected the tree in Figure 5.1 constructed with the parameter MNIL 7.



ACC 71.86%; CCPE 0.5879; Kappa 0.5497; CPX 8

Figure 5.1: The initial tree constructed from the 60 higher education attributes.

Step 6: Modify the initial tree.

Remove attributes. A removed attributes graph is presented in Figure 5.2. The numbers within the credibility indicators divided by the vertical bar represent: ACC, CCPE, Kappa and q_{Δ}. The numbers in brackets that follow a quality estimate serve for tracking the changes in quality, where the first number represents the difference in quality between the current indicator and the initial tree, while the second number represents the difference between the current indicator and the superordinate indicator. Multiple quality measures are presented in the graph to provide enough evidence for the user to make an informed decision when comparing several models. For example, a fall in all the quality measures indicates a relation of higher credibility than the fall is some of the quality measures. The indicators are written in blue. In contrast, the indicators that the HMDM method considers as less-credible (the attributes that do not modify either quality or structure of the model) are written in strikethrough.

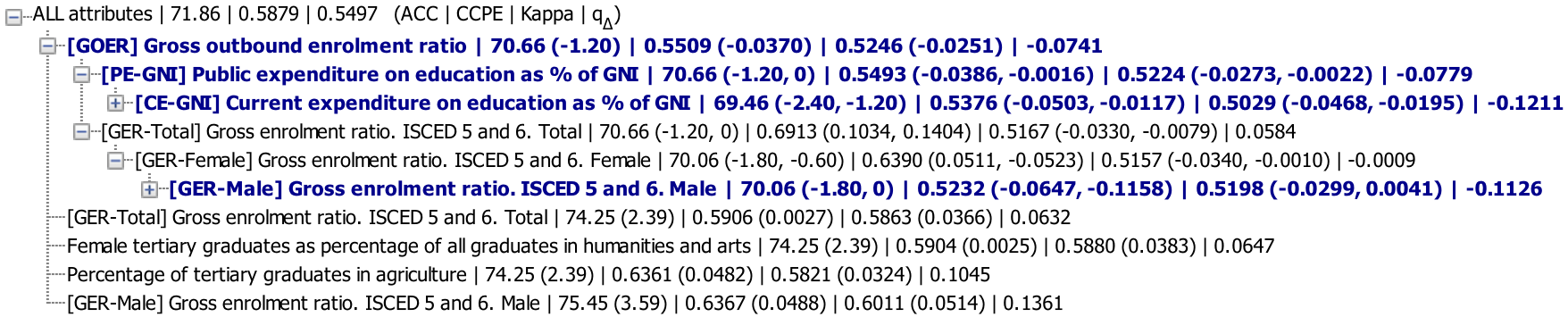


Figure 5.2: The removed attributes graph constructed from the 60 higher education attributes.

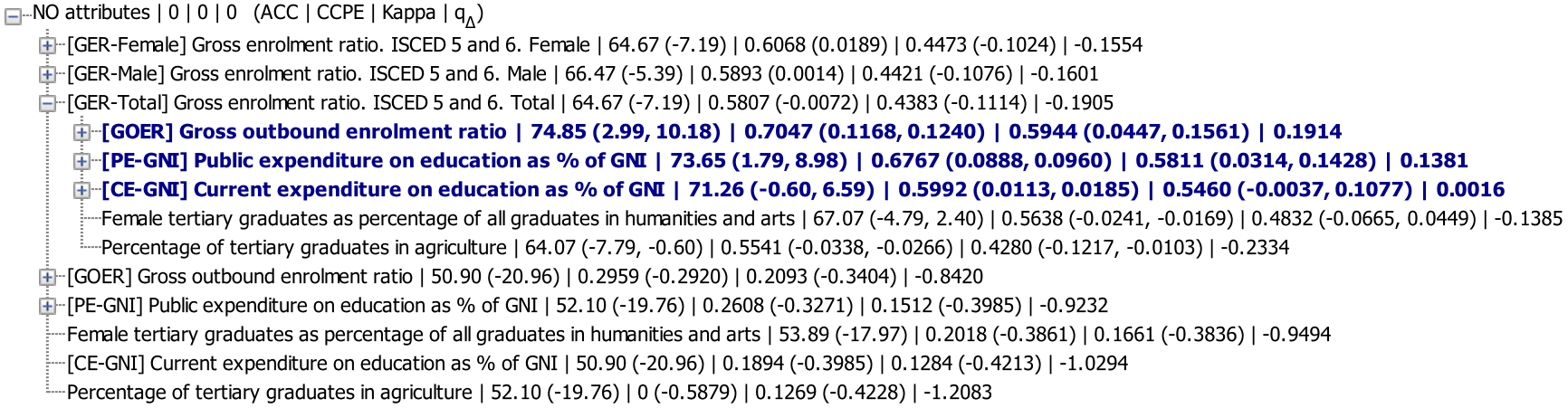


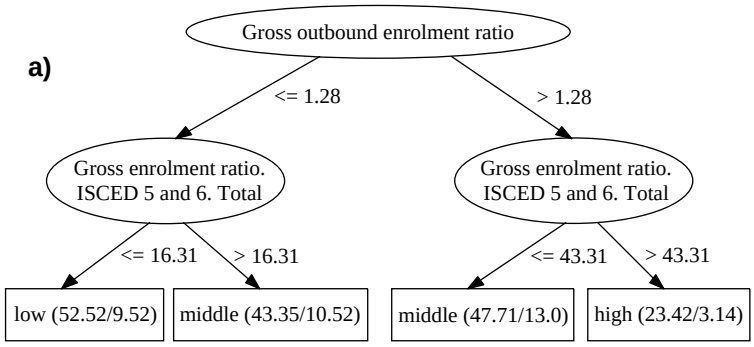
Figure 5.3: The added attributes graph constructed from the higher education attributes.

The graph reveals three findings. First, the mobility of students is important for a country's welfare. "Gross outbound enrolment ratio" (GOER), which indicates the level of students' participation in the education programs conducted by foreign higher education institutions, emerged as the most credible at the first level of the graph. When removed, all three quality estimates fell, resulting in a q_Δ of -0.0741. Second, the level of investment in all levels of education emerged as credible. When "Public expenditure on education as % of GNI" (PE-GNI) is removed in addition to GOER, two quality estimates further decreased (q_Δ of -0.0779). The PE-GNI represents a proportion of country's wealth that has been spent on education during a given year. The PE-GNI is substituted by the "Current expenditure on education as % of GNI" (CE-GNI) attribute, which represents the same semantic category. Generally, PE-GNI covers both current and capital expenditures. While current expenditures include expenditures for goods and services (e.g., staff salaries, teaching materials), capital expenditures include expenditures for assets that last longer than one year (e.g., for construction of buildings). Since current expenditures form the majority of the public expenditures, these two attributes are equally treated. Furthermore, when PE-GNI is removed alone, the fall in q_Δ is negligible (0.0038), while when both attributes are removed the fall in q_Δ is considerably higher (0.047). Therefore, the two attributes form a PE-GNI||CE-GNI relation, which is also confirmed by the negative interaction of -0.049 computed between the two attributes and the class (note that in this thesis the class attribute is added to all subsets for which the interaction is computed, even though it is not explicitly stated). Third, the level of participation in higher education is important for welfare. The three "Gross enrolment ratio. ISCED 5 and 6" (GER) attributes – Total, Male and Female substitute each other in the same way as PE-GNI and CE-GNI, forming the GER relation (GER-Total||GER-Male||GER-Female) – q_Δ of -0.1126. The other attributes appear to be less credible.

Add attributes. The added attributes graph in Figure 5.3 reveals two additional findings. First, the best result is obtained by combining the GOER with the GER attributes. The highest q_Δ of 0.1914 is observed when the tree is constructed from the GOER and GER-Total attributes (see Figure 5.4a). The trees of similar structure and quality are obtained by substituting the GER-Total with any of the GER-Female or GER-Male attributes; therefore, the combination will be denoted as GOER&GER. The tree in Figure 5.4a represents credible relations, indicating that for better welfare it is important to stimulate participation in higher education and to improve the student exchange programs, especially for those students that leave the country to study abroad. Looking back to the removed attributes graph in Figure 5.2, a considerable fall in quality (q_Δ of -0.1126) is observed when the GOER and the GER attributes are removed in combination, providing additional evidence that supports the GOER&GER combination as the first level of credibility. In addition, the positive interaction supports the relation: GOER&GER-Total (0.0193), GOER&GER-Male (0.0095) and GOER&GER-Female (0.0037). Second, the PE-GNI||CE-GNI is supported by two trees of which one, constructed from the GER-Total and PE-GNI (q_Δ of 0.1381), is presented in Figure 5.4b. The other, constructed from the GER-Total and CE-GNI attributes (q_Δ of 0.0016), differs from the first tree only in the left subtree, where the PE-GNI subtree in Figure 5.4b is substituted by the CE-GNI subtree, dividing "low" and "middle" countries in the same manner. The two trees represent the credible relations, indicating that to improve the welfare of developing countries ("low" and "middle") it is important to increase the general level of investment in education.

Step 7: Integrate conclusions.

Several tens of additionally constructed trees confirmed the first impression that the most important measure to improve welfare is to stimulate participation in higher ed-



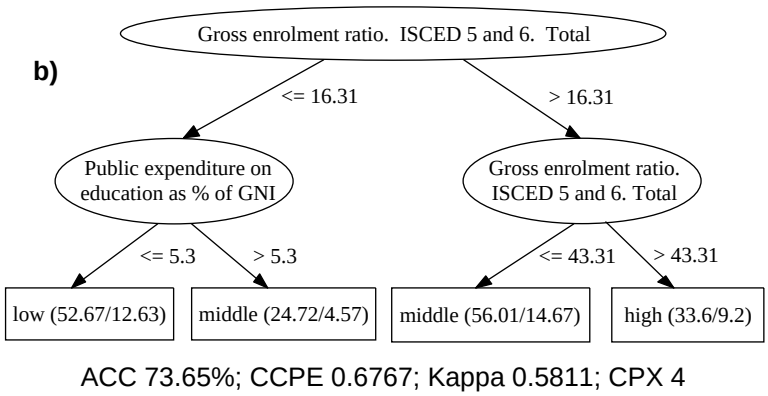


Figure 5.4: Two credible trees constructed from: a) the GER-Total and GOER; b) the GER-Total and PE-GNI.

ucation and to improve the student exchange programs, especially for those students that leave the country to study abroad (GOER&GER). In addition, developing countries should increase their levels of investment for all levels of education (PE-GNI||CE-GNI).

5.1.2 Decision Trees Constructed from the Modified Attribute Set

Step 2: Modify attribute set.

Nine attributes are constructed based on the observations, e.g., "GER-Total + GOER" is constructed by summing GER-Total and GOER (see Appendix A for the complete list of the constructed attributes). Twenty attributes differentiating between the status of female and male students are removed (the better status of females in higher education is more a consequence of better economic welfare, than a possible cause).

Step 3,4: The same setup is used as in Subection 5.1.1.

Step 5: The INITIAL_DM resulted in 16 trees, from which we selected the tree in Figure 5.5 constructed with the parameter MNIL 14.

Step 6: Modify the initial tree.

Remove attributes. A removed attributes graph in Figure 5.6 is constructed under the assumption that the most credible attribute will always appear in the root of the tree; therefore, the indicators within the graph (except the last two) were constructed by iteratively removing the root attribute until all the interesting attributes were removed. The graph confirms the conclusions made with the original attribute set and brings some additional insights. First, the credibility of the GOER&GER relation is supported by the

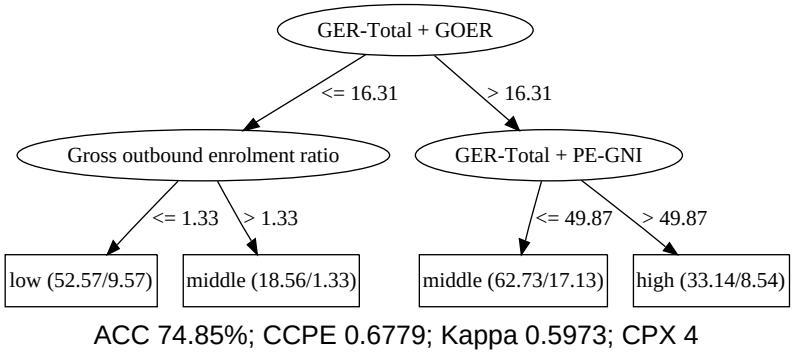


Figure 5.5: The initial tree constructed from the modified higher education attribute set.

highest fall in quality at the first level of the graph (q_Δ of -0.0828) when the constructed attribute "GER-Total + GOER" is removed. In addition, when another constructed attribute "GER-Total + PE-GNI" is removed in addition to "GER-Total + GOER", the tree equal to the one in Figure 5.4a is constructed, which is of quality similar to the quality of the initial tree (q_Δ of 0.0135). After removing the two attributes from the constructed tree the quality, further down the graph, considerably falls (q_Δ of -0.2978). Second, an additional weight is added to the level of participation in higher education, denoted by the GER. When GER-Total is removed, the attribute "Tertiary students per 100,000 inhabitants" (TERT-STUD) takes its place as a root of the tree. The TERT-STUD also represents the level of participation in higher education, but expressed in different quantity. Considering that the TERT-STUD attribute represents the same semantic category as the GER attributes, the attribute can be added to the GOER&GER relation, resulting in the GOER&(GER||TERT-STUD) relation. The updated relation is further supported by: a) a considerable fall in quality (q_Δ of -0.8352) when all the attributes containing GOER, GER and TERT-STUD are removed; b) negative interaction computed for each relation containing one of the GER attributes and the TERT-STUD attribute, and positive interaction for the GOER&TERT-STUD relation. Third, the relation PE-GNI||CE-GNI is established as the second level of credibility relation: the removal of the constructed attribute "GER-Total + PE-GNI" results in a quality increase (q_Δ of 0.0135), while the removal of the constructed attribute "GER-Total + CE-GNI" results in a quality decrease, which represents the conflicting evidence. In addition, the attribute "Public expenditure on education as % of GDP (gross domestic product)" (PE-GDP) emerged as credible, which represents the same semantic category as PE-GNI and CE-GNI. When the attribute "GER-Total + CE-GNI" is removed, the PE-GDP takes its role in the tree. Based on the evidence, the PE-GNI||CE-GNI is updated with the PE-GDP attribute, consequently forming the relation PE-GNI||CE-GNI||PE-GDP. The new relation is supported by the negative interactions for PE-GNI||PE-GDP (-0.0794) and for CE-GNI||PE-GDP (-0.0561).

Add attributes. The added attributes graph in Figure 5.7 is constructed interactively: first, by adding all the attributes; then, by iteratively expanding the most promising indicators, again with all the available attributes. In total, there are four important findings. First, the initial tree in Figure 5.5 is the best observed tree (q_Δ of 0.0507). The tree generally supports the GOER&(GER||TERT-STUD) relation. The right subtree, at first sight, seems to be providing supporting evidence for the PE-GNI||CE-GNI||PE-GDP, indicating that the PE-GNI is capable of differentiating between the "high" and "middle" countries. However, additional analyses refuted the claim by showing that the right subtree was formed primarily because of the important role of the GER-Total within the "GER-Total + PE-GNI" attribute. Second, the GOER&(GER||TERT-STUD) is ad-

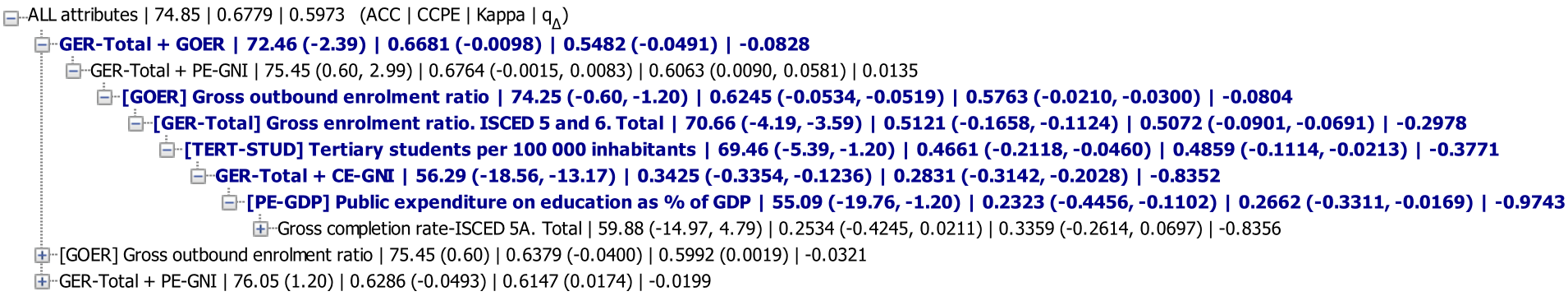


Figure 5.6: The removed attributes graph constructed from the modified higher education attribute set.

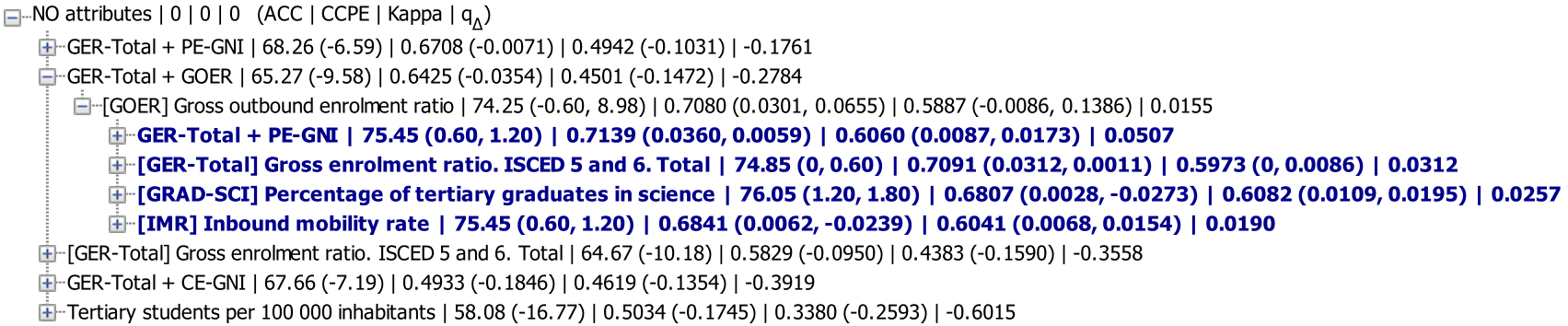


Figure 5.7: The added attributes graph constructed from the modified higher education attribute set.

ditionally supported by the tree constructed from the "GER-Total + GOER", GOER and GER-Total attributes (q_Δ of 0.0312). Third, "Percentage of tertiary graduates in science" (GRAD-SCI) emerged as a second level of credibility, differentiating "high" from "middle" countries according to the higher percentage of tertiary graduates in science (q_Δ of 0.0257). Fourth, "Inbound mobility rate" (IMR) also emerged as the second level of credibility. The IMR differentiates "high" from "middle" countries according to the larger number of foreign students (q_Δ of 0.0190).

Step 7: Integrate conclusions.

As in previous cases, only a minor portion of all the examined trees is presented. The analyses with the modified attribute set confirmed the conclusions made with the original attribute set. In these tests, two additional second level of credibility relations were discovered. The attributes indicate that for "middle" countries to improve their welfare it is important to attract more foreign students and to increase the number of graduates in science programs (life and physical sciences, mathematics and statics, computing – software development) (UNESCO, 2006).

5.1.3 Regression Trees Constructed from the Modified Attribute Set

Step 1: The data set from Subsection 5.1.1 is used, except that the class is transformed into a numerical representation: low is encoded as 1, middle as 2 and high as 3. Within the preliminary research (Vidulin and Gams, 2011) we experimented with the class stated in US$; however, the resulting regression trees were of poor quality (CC of 0.52).

Step 2: 20 attributes differentiating female from male students are removed.

Step 3: Regression trees are constructed with the M5P (Quinlan, 1992) algorithm from Weka.

Step 4: The MNIL parameter of M5P is selected with values ranging from the default 4 to 15. The MNR is set to 2 and remove duplicate models to "on".

Step 5: The INITIAL_DM returned three trees, from which we selected the tree in Figure 5.8[1] constructed with the parameter MNIL 15.

Step 6: Modify the initial tree.

Remove attributes. A removed attributes graph in Figure 5.9 reveals three findings. First, the credibility of GOER&(GER||TERT-STUD) is supported by the highest observed fall in quality when the three attributes are removed (q_Δ of -0.3708). When TERT-STUD is removed, GER-Total takes its role, confirming the GER||TERT-STUD part of the relation. Second, the attribute "Public expenditure per pupil as a % of GDP per capita" (PE-PUP-GDP-PC) emerged as credible (q_Δ of -0.0131), which is semantically similar to the attributes within the PE-GNI||CE-GNI||PE-GDP relation. However, two pieces of evidence indicate that the PE-PUP-GDP-PC attribute does not belong to the PE-GNI||CE-GNI||PE-GDP relation: a) when removed none of the attributes from the relation took its role in the tree; b) interactions between the PE-PUP-GDP-PC attribute and each of the attributes from the relation are positive. Since the PE-PUP-GDP-PC has not been supported by the graphs observed up to this point, it belongs to the second level of credibility. Third, "Percentage of tertiary graduates in education" is less-credible,

[1]The first number within a leaf represents an average of target attribute values for those examples that reached the leaf. The first number in brackets represents the number of examples within the leaf, while the second number represents the deviation around the average value, computed by dividing the root-mean-squared error by the global absolute deviation. The values within the leaves are directly comparable with the classes: low < 1.5, middle $[1.5, 2.5\rangle$ and high ≥ 2.5.

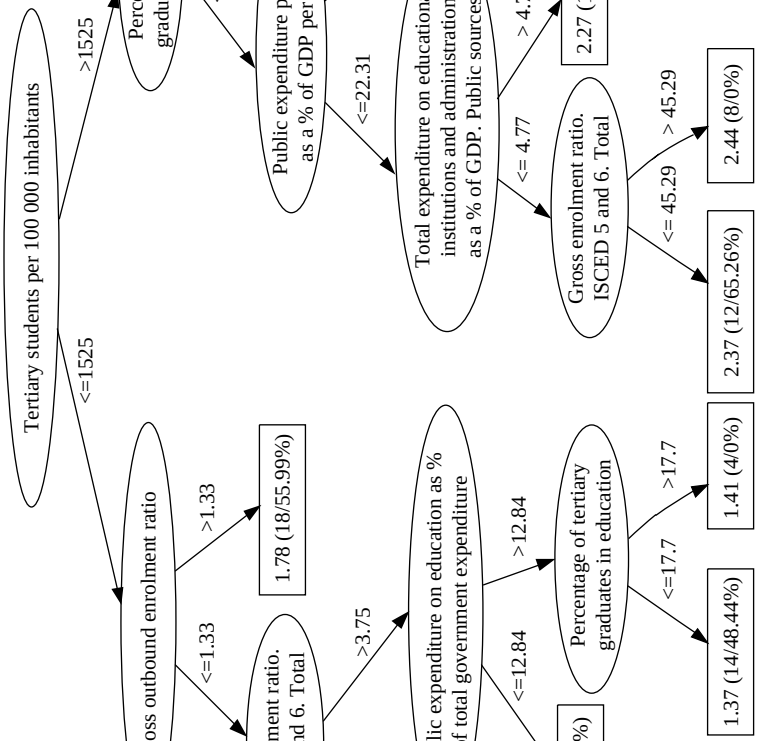


Figure 5.8: The initial tree constructed from the modified higher education attribute set with the regression trees.

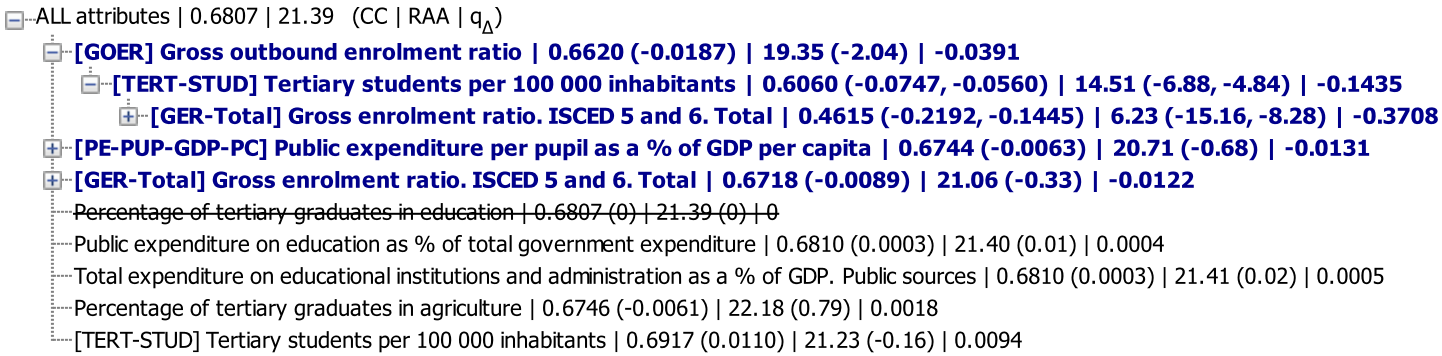


Figure 5.9: The removed attributes graph constructed from the modified higher education attribute set with the regression trees.

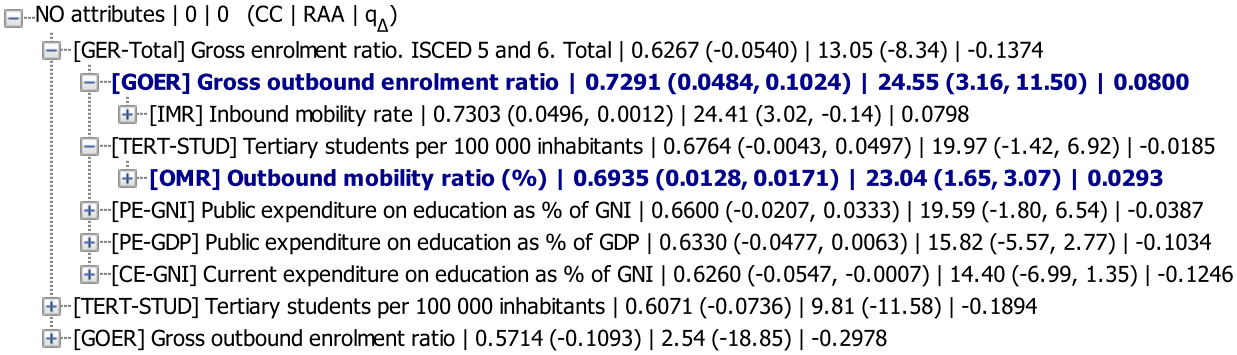
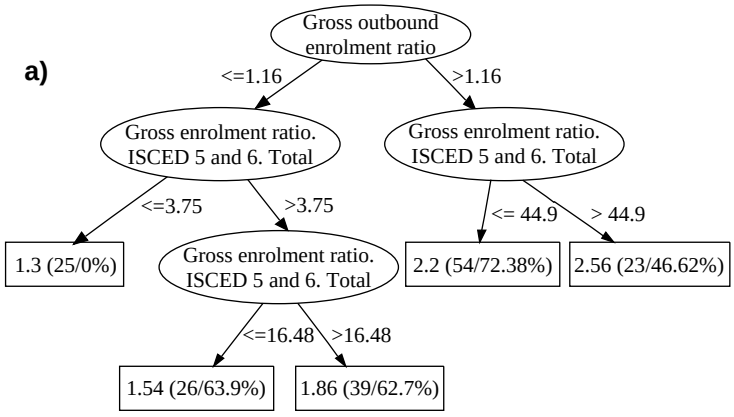


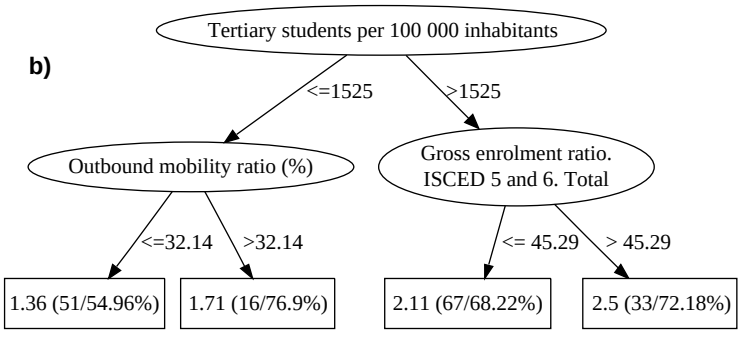
Figure 5.10: The added attributes graph constructed from the modified higher education attribute set with the regression trees.

since the quality remains the same when the attribute is removed. From the ten trees constructed on cross-validation folds from the complete attribute set, the less-credible attribute did not appear in any, showing that the attribute was included in the initial tree based on the specific data sample.

Add attributes. The added attributes graph in Figure 5.10 is constructed with the same approach as the added attributes graph in Subsection 5.1.2. At the first level of the graph, only those attributes with a CC higher than 0.5 were retained. The graph reveals two findings. First, the GOER&(GER||TERT-STUD) is supported by: a) the appearance of these three attributes at the first level of the graph, while the other attributes did not exceed the 0.5 limit; b) the credible relations within the best observed tree (q_Δ of 0.08) constructed from GOER and GER (see Figure 5.11a), similar to those within the tree in Figure 5.4a; c) the credible relations within the high-quality tree (q_Δ of 0.0293) constructed from GER, TERT-STUD and "Outbound mobility ratio (%)" (OMR) attributes (see Figure 5.11b) – the tree represents semantically similar relations as the tree in Figure 5.11a, since the OMR substitutes the GOER, which belongs to the same semantic category. The result is the (GOER||OMR)&(GER||TERT-STUD) relation. Second, the credibility of IMR is supported by the tree of quality higher than the initial tree (q_Δ of 0.0798). However, the quality of superordinate tree, which is constructed from GER-Total and GOER, is not improved by adding the IMR attribute. Therefore, the IMR is confirmed as the second level of credibility.



CC 0.7291; RAA 24.55%; CPX 5



CC 0.6935; RAA 23.04%; CPX 4

Figure 5.11: The credible trees constructed from: a) GOER and GER; b) TERT-STUD, GER and OMR.

Step 7: Integrate conclusions.

The most important measure is again to stimulate participation in higher education and to improve the student exchange programs, especially for those students that study abroad (GOER||OMR)&(GER||TERT-STUD). Second, for the developing countries to improve their welfare, it is also important to increase the level of investment in all levels of education (PE-GNI||CE-GNI||PE-GDP, PE-PUP-GDP-PC). Third, to further improve the welfare of the "middle" countries, it is important to increase the number of graduates in the science programs (GRAD-SCI), as well as to attract foreign students (IMR). Note that these relations are not harmful for other countries; they are just not so important anymore. Some other relations like "Percentage of tertiary graduates in education" emerged during the initial DM, but additional analyses revealed that they belong to the third level of credibility and are not discussed further.

Several of the presented relations are recognized within the related work, showing the ability of the HMDM method to find credible relations. The importance of the level of participation in higher education was recognized by Keller (2006). Furthermore, the importance of the investment in education was acknowledged by Gylfason (2001), while the importance of participation in science higher education programs was noted by Varsakelis (2006). With the help of the HMDM method we further discovered that a higher mobility of students is also very important for better economic welfare. This relation was not directly discussed in the related work (Gylfason, 2001; Keller, 2006; Varsakelis, 2006). In addition, our method provides a classification scheme that differentiates not only credible from less-credible relations, but also indicates how credible the discovered relations are and presents the relations in a human-readable form.

In summary, this section presented an application of the HMDM method for the analysis of higher education data. Some of the credible relations were already established in the literature, while some of them were not directly discussed. The credibility of discovered relations is further supported by the evaluation presented in Section 5.3, which shows that the relations represent non-random patterns. In the following section, the HMDM method will be applied in the R&D domain.

5.2 The Impact of the R&D Sector on Economic Welfare

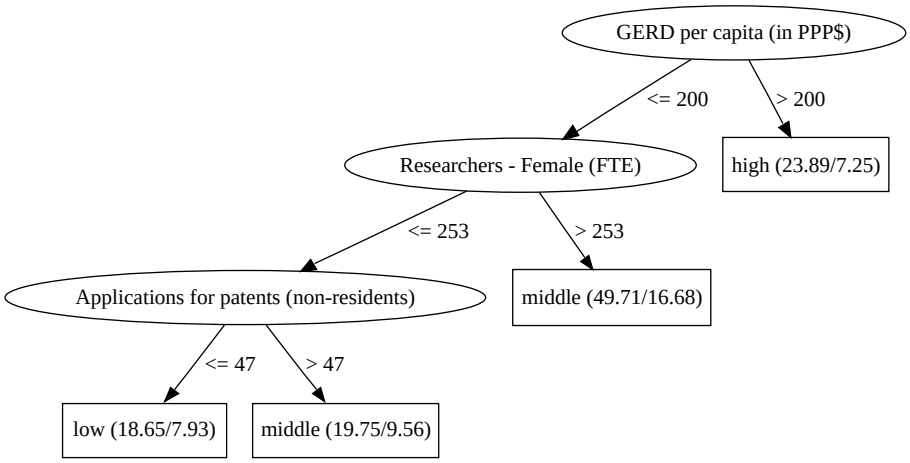
We collected the data representing the R&D sector from two statistical databases provided by the UNESCO Institute for Statistics (http://www.uis.unesco.org) and WIPO (http://www.wipo.int). The economic welfare is represented by two "GNI per capita" attributes, described in Section 5.1. The data set is available at http://dis.ijs.si/Vedrana/economic-analysis.htm, while the attributes are described in Appendix B.

The following subsections present three HMDM analyses of the R&D data, performed by constructing: first, decision trees from the original attribute set (Subsection 5.2.1); second, decision trees from the modified attribute set (Subsection 5.2.2); and third, regression trees from the modified attribute set (Subsection 5.2.3).

5.2.1 Decision Trees Constructed from the R&D Data

Step 1: The data set is composed of 48 attributes describing the R&D sector, and 167 examples representing countries. The class is discrete "GNI per capita".
Step 3,4: The same setup is used as in Subsection 5.1.1.

Step 5: The INITIAL_DM returned 20 trees, from which we selected the tree in Figure 5.12 constructed with the parameters MNIL 12 and REP.



ACC 63.47%; CCPE 0.4438; Kappa 0.378; CPX 4

Figure 5.12: The initial tree constructed from the 48 R&D attributes.

Step 6: Modify the initial tree.

Remove attributes. A removed attributes graph is presented in Figure 5.13. The graph reveals five findings. First, the level of investment in R&D is important for a country's welfare, since "GERD per capita (in PPP$)" (GERD-PC) is the most credible attribute at the first level of the graph. GERD stands for Gross Domestic Expenditure on R&D, denoting the expenditure on R&D performed on the national territory during a year (OECD, 2002). PPP$ stands for purchasing power parity in American dollars. A statement of GERD in PPP$ allows for fair comparisons between the countries. When GERD-PC was removed, two out of three quality estimates fell, resulting in q_Δ of -0.0758. An increase in CCPE indicates the appearance of a redundant attribute "GERD as % of GDP" (GERD-GDP) in a root of the tree, which was constructed after the GERD-PC has been removed. The removal of both attributes caused a fall in all three quality estimates (q_Δ of -0.2534), which confirmed the relation GERD-PC||GERD-GDP. In addition, the negative interaction of -0.3758 between the two GERD attributes provides further support. Second, the number of researchers emerged as credible. After the removal of both GERD attributes, "Researchers per million inhabitants (HC = head count)" (RES-HC) appeared as the next important attribute. When removed, the q_Δ further decreased, resulting in a q_Δ of -0.2576. Further, the redundant attribute "Researchers per million inhabitants (FTE = full-time equivalent)" (RES-FTE) took its role as a root of the tree, but this time, when RES-FTE was removed, the q_Δ did not further decrease and none of the redundant attributes took its place. Therefore, a redundant relation RES-HC||RES-FTE was established, which belongs to the second level of credibility. The redundancy is confirmed by the negative interaction of -0.2976. Third, the number of applications for patents, where the first applicant is a non-resident of a country, emerged as credible. On the fourth level of the graph, the removal of the attribute "Application for patents (non-residents)" (APP-NON-RES) caused a further fall in all three quality estimates (q_Δ of -0.349). Fourth, an export of high-technology products and services, which are the result of high-intensity R&D activities, is important for the welfare. "High-technology exports (% of manufactured exports)" (HI-TECH) is the only attribute, in addition to APP-NON-RES, which caused a further fall in q_Δ in comparison to the superordinate indicator. Fifth, the number of female researchers is less-credible. When removed, the

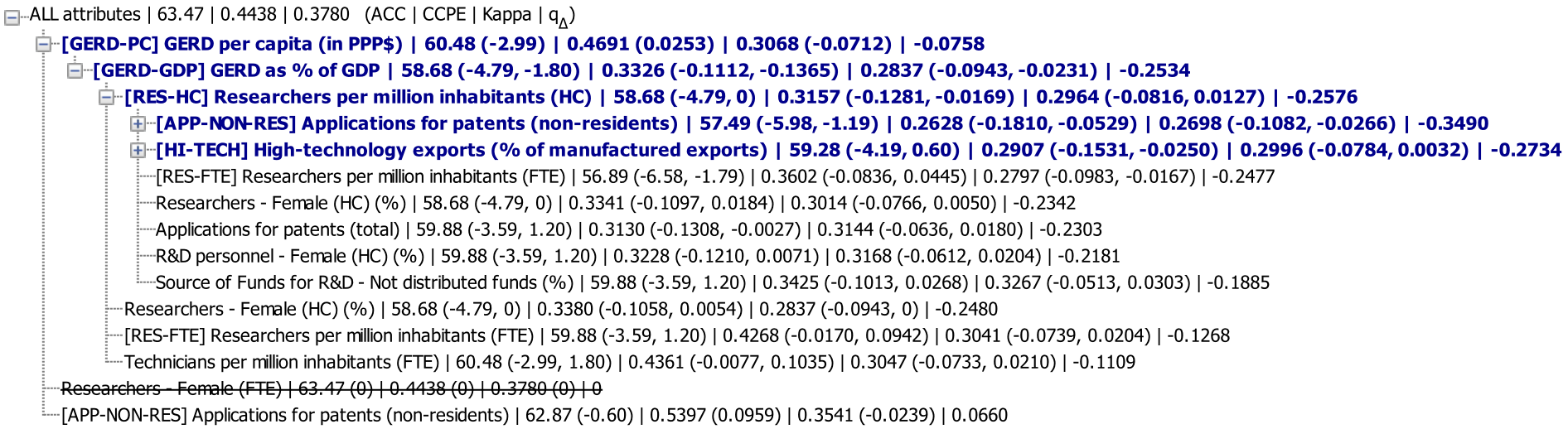


Figure 5.13: The removed attributes graph constructed from the 48 R&D attributes.

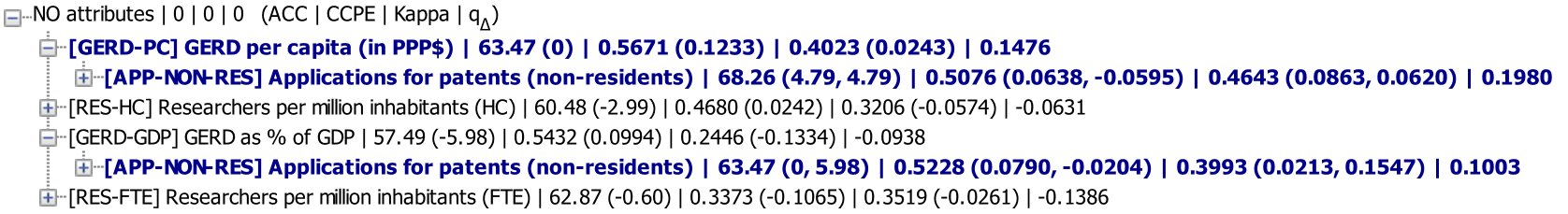
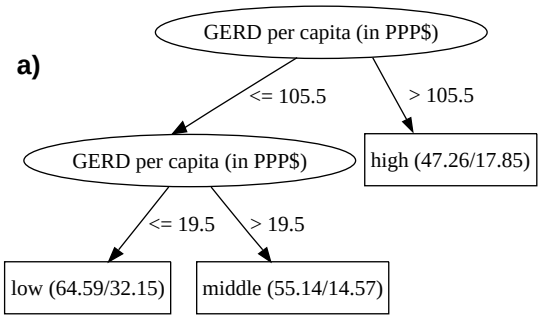
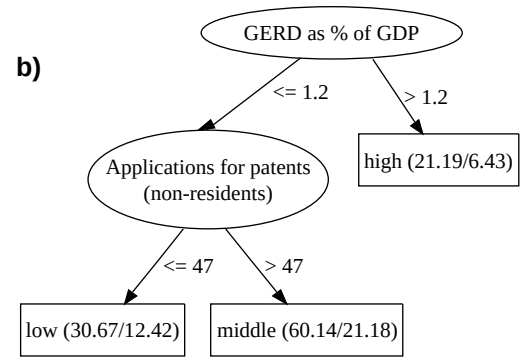


Figure 5.14: The added attributes graph constructed from the 48 R&D attributes.

attribute "Researchers – Female (FTE)" did not cause any change in quality. From the ten trees constructed on cross-validation folds from the complete attribute set, the less-credible attribute did not appear in any, showing that the attribute was included in the initial tree based on the specific data sample.



ACC 63.47%; CCPE 0.5671; Kappa 0.4023; CPX 3



ACC 63.47%; CCPE 0.5228; Kappa 0.3993; CPX 3

Figure 5.15: The credible trees constructed from: a) GERD-PC; b) GERD-GDP and APP-NON-RES.

Add attributes. The added attributes graph presented in Figure 5.14 reveals three findings. First, the graph confirms that the level of investment in R&D is indeed important for a country's welfare. The tree constructed only from the GERD-PC (see Figure 5.15a) is of a quality higher than the initial tree (q_Δ of 0.1476), at the same time containing credible relations, which indicate that a higher level of investment in R&D leads to better welfare. Second, APP-NON-RES is the second level of credibility attribute. It appeared in two combinations, GERD-PC&APP-NON-RES and GERD-GDP&APP-NON-RES, where both trees were of a quality higher than the initial tree (q_Δ of 0.198 and 0.1003, respectively). Both trees contain credible relations, showing that a higher level of investment in R&D leads to better welfare, while "low" countries should increase the number of applications for patents submitted by non-residents to improve their welfare (the second tree is presented in Figure 5.15b). However, the interaction for GERD-PC&APP-NON-RES is negative (-0.0521), while for GERD-GDP&APP-NON-RES it is positive (0.01), which is contradictory and indicates the second level of credibility relations. Considering: a) the strong evidence that GERD-PC||GERD-GDP belongs to the first level of credibility, and b) that APP-NON-RES describes a subgroup within the data (the difference between "low" and "middle" countries) and needs the support of another strong attribute to describe the rest of the examples, we extracted APP-NON-RES as a single important attribute belonging to the second level of credibility group. Third, the RES-HC||RES-FTE relation is confirmed to be the second level of credibility relation. At the

first level of the graph, the two RES attributes are the closest in quality to the two GERD attributes. At the same time, the trees constructed from other attributes at the first level of the graph were not of acceptable quality and are, consequently, not presented in the graph. The two trees, constructed from each of the RES attributes, both contain credible relations, where a larger number of researchers leads to better welfare. However, when added in combination with the first level of credibility attributes, the RES attributes are not strong enough to be included in the tree.

Step 7: Integrate conclusions.

The analysis showed that the most important factor to improve welfare is to increase the level of investment in the R&D sector (GERD-PC||GERD-GDP). In addition, the countries should increase the number of researchers (RES-HC||RES-FTE), while developing countries should further consider the help of foreign experts in order to increase the number of patents (APP-NON-RES). Some evidence was observed that the amount of exported products and services, which are the results of R&D activities (HI-TECH), is important for a country's welfare, however, further analyses are needed to confirm its credibility. Some other attributes appeared in the analysis, such as the number of female researchers, however, they belong to the third level of credibility.

5.2.2 Decision Trees Constructed from the Modified Attribute Set

Step 2: Modify attribute set.

A total of 19 attributes are constructed based on the observations. For example, an attribute "Sector investing the most in R&D" is constructed by finding the maximum between six "Source of funds for R&D" attributes. Accordingly, the attribute takes one of six values: business enterprise, government, higher education, private non-profit, abroad and N/A (not known or not distributed funds). The constructed attributes are described in detail in Appendix B.

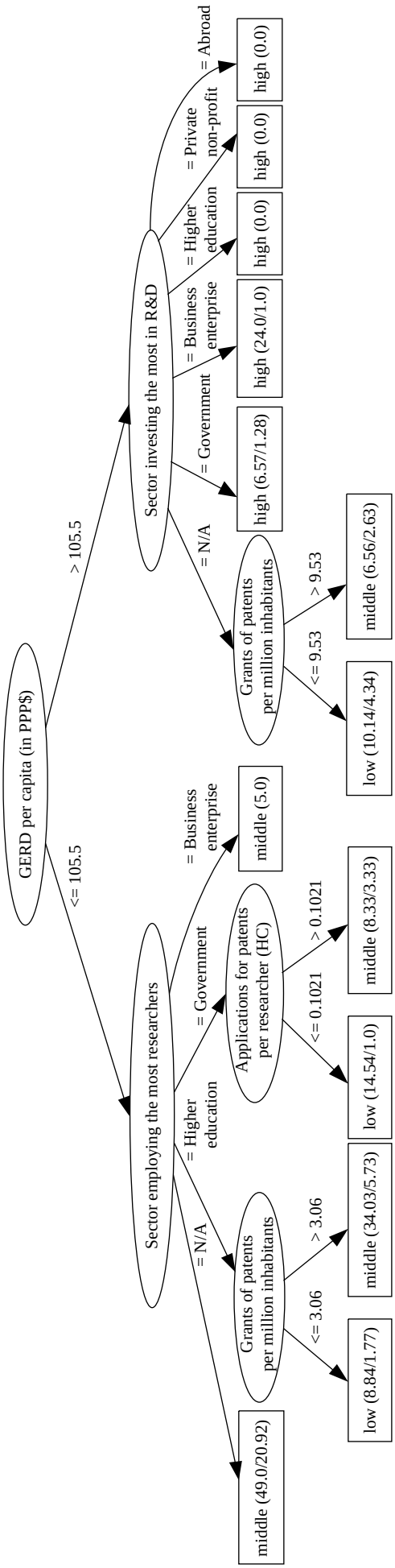
Step 3,4: The same setup is used as in Subsection 5.2.1.

Step 5: The INITIAL_DM resulted in 18 trees, from which we selected the tree in Figure 5.16 constructed with the parameter MNIL 2.

Step 6: Modify the initial tree.

Remove attributes. A removed attributes graph presented in Figure 5.17 reveals the following findings.

First, the graph confirms the important role of investment in the R&D sector, since GERD-PC appeared as the most credible attribute at the first level of the graph (q_Δ of -0.0841). When GERD-PC was removed, the constructed attribute "GERD as % of GNI" (GERD-GNI) took its role as a root of the tree. The GERD-GNI represents the same semantic category as the GERD-PC, only expressed in different quantity. The removal of both GERD attributes resulted in a further fall in quality (q_Δ of -0.0886) caused by changes in ACC and Kappa. The difference in CCPE, however, remained positive due to appearance of GERD-GDP within the newly constructed tree. The GERD-GDP attribute does not have the role of the root, however, it describes the semantically similar relations as the first two GERDs. When we removed all three GERD attributes, all quality measures decreased (q_Δ of -0.235). Therefore, we established the redundant relation GERD-PC||GERD-GNI||GERD-GDP, which was also supported by the negative interactions: GERD-PC||GERD-GNI of -0.403; GERD-PC||GERD-GDP -0.3758; GERD-GNI||GERD-GDP of -0.4452.



ACC 69.46%; CCPE 0.374; Kappa 0.4863; CPX 13

Figure 5.16: The initial tree constructed from the modified R&D attribute set.

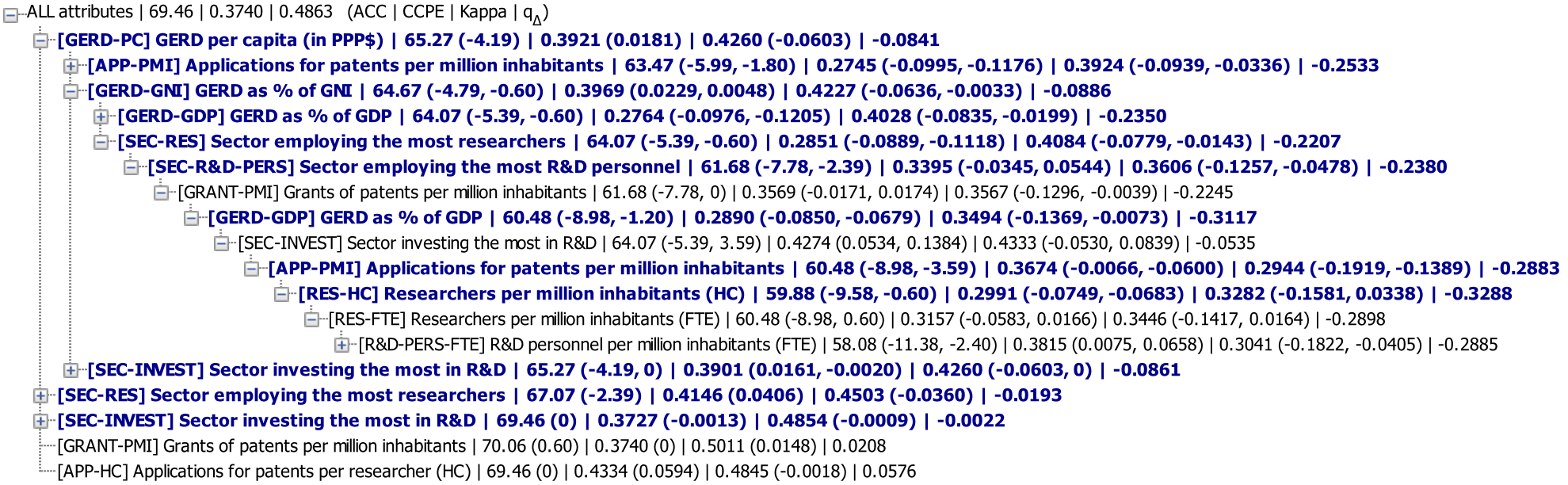


Figure 5.17: The removed attributes graph constructed from the modified R&D attribute set.

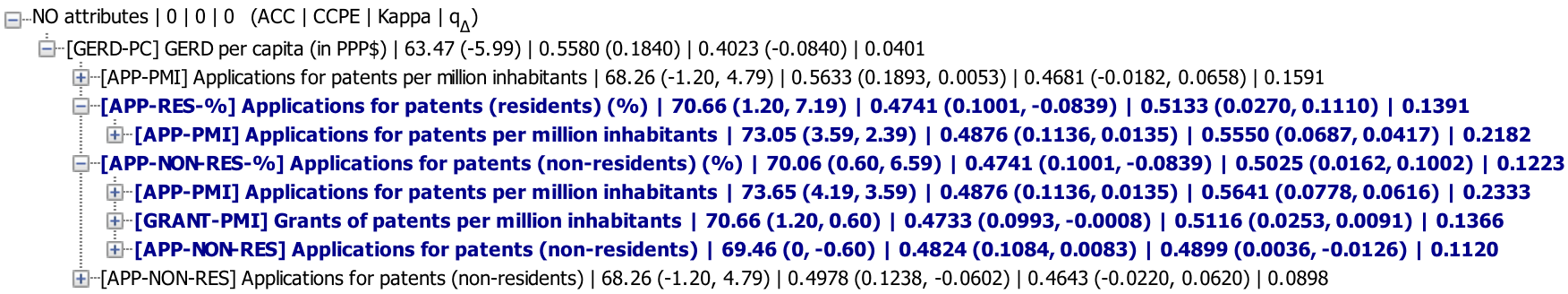


Figure 5.18: The added attributes graph constructed from the modified R&D attribute set.

Second, the graph shows that it is not only important how many researchers are employed, but also in which sector they are employed. After the removal of GERD-PC and GERD-GNI, we removed a sequence of attributes in the order in which they continued to appear in the root of the tree. The next root attribute was "Sector employing the most researchers" (SEC-RES), which indicates the sector that performs the most R&D activities in a country, with the emphasis on work done by researchers (excluding technicians and other supporting staff). When removed, the quality further decreased (q_Δ of -0.2207), while its role was taken by the semantically similar attribute "Sector employing the most R&D personnel" (SEC-R&D-PERS). In comparison to the SEC-RES, the SEC-R&D-PERS also accounts for the work done by technicians and other supporting staff. The removal of the SEC-R&D-PERS caused a further fall in quality (q_Δ of -0.238), indicating the redundant relation SEC-RES||SEC-R&D-PERS. The redundancy was confirmed by a negative interaction of -0.174.

Third, the constructed attribute "Sector investing the most in R&D" (SEC-INVEST) emerged as the second level of credibility attribute, indicating that for welfare it is important which sector invests the most in R&D activities. This attribute appeared three times in the graph. At the first level of the graph, the removal of the attribute resulted in a small decrease in quality (q_Δ of -0.0022). At the second level of the graph, when removed together with the GERD-PC attribute, the quality further slightly decreased (for 0.002 in comparison to the superordinate tree). However, within the sequence of root attributes, the removal of SEC-INVEST did not result in a further fall in quality. On the contrary, the q_Δ increased by 0.2582.

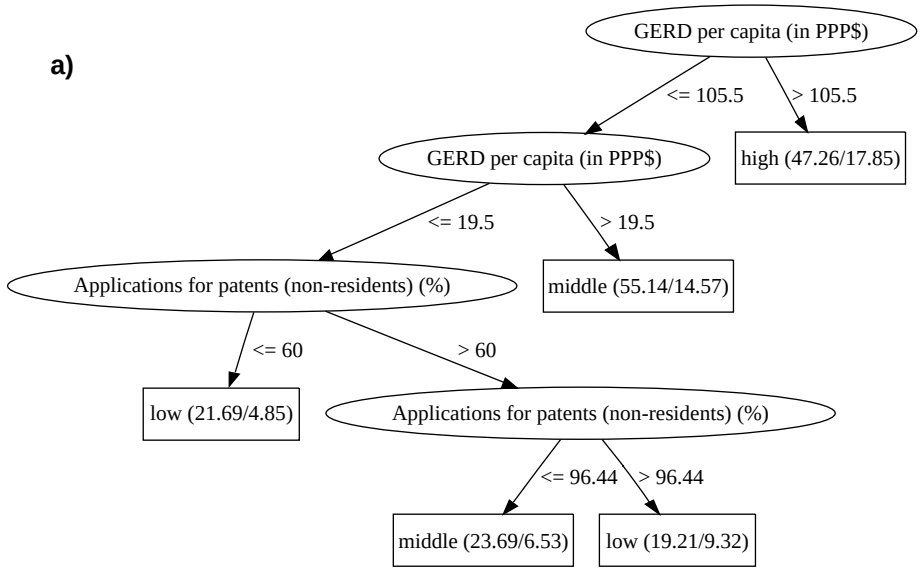
Fourth, the number of applications for patents was confirmed as credible, however, this time as the total number of applications for patents. The constructed attribute "Applications for patents per million inhabitants" (APP-PMI) appeared two times in the graph, both times further decreasing the quality (q_Δ of -0.2533 and -0.2883).

Finally, RES-HC||RES-FTE confirmed as the second level of credibility relation. Within the queue of root attributes, the removal of RES-HC first reduced the quality (q_Δ of -0.3288), while the removal of RES-FTE then increased the quality (q_Δ of -0.2898).

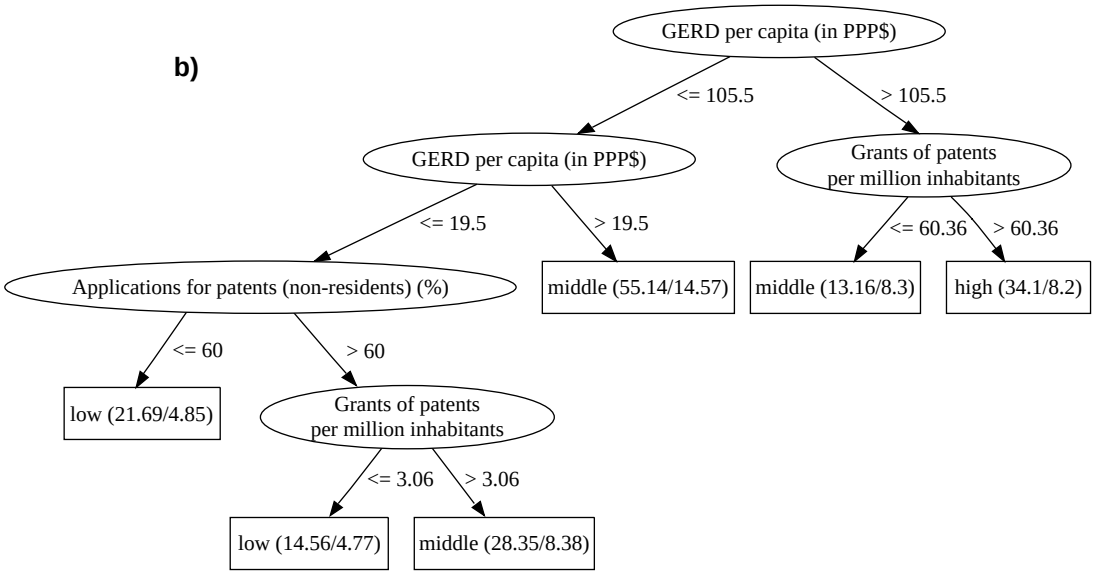
Add attributes. The added attributes graph presented in Figure 5.18 reveals multiple evidence indicating that for improving a country's welfare, it is important to increase the level of investment in R&D activities and to stimulate the researchers to innovate, which is reflected in more patents.

First, the graph indicates the combination GERD-PC&APP-PMI as the second level of credibility relation, which is partially supported by and partially opposed by the following findings: a) the tree constructed from the two attributes is of better quality than the initial tree (q_Δ of 0.1591) due to the increase in CCPE, but it is not supported by the increase in the majority of the quality measures; b) the quality further improves when the two attributes are added in combination with other attributes, e.g., with "Applications for patents (residents)(%)" (APP-RES-%) (q_Δ of 0.2182) and "Applications for patents (non-residents)(%)" (APP-NON-RES-%) (q_Δ of 0.2333); c) both attributes form credible relations, which differentiate between all three types of countries; and d) the interaction is negative (-0.3376), while it should be positive to support the combination.

Second, the graph supports and additionally explains the second level of credibility relation APP-NON-RES. Two constructed attributes, which are semantically similar to APP-NON-RES, emerged as credible: APP-NON-RES-% and APP-RES-%. The two attributes together represent a distribution of patents among the residents and non-residence of a country, where APP-NON-RES was used for constructing both of them. Two trees constructed from: a) GERD-PC and APP-RES-% (q_Δ of 0.1391), and b) GERD-PC and APP-NON-RES-% (q_Δ of 0.1223), both indicate that in "low" countries non-residents



ACC 70.06%; CCPE 0.4741; Kappa 0.5025; CPX 5



ACC 70.66%; CCPE 0.4733; Kappa 0.5116; CPX 6

Figure 5.19: The credible trees constructed from: a) GERD-PC and APP-NON-RES-%; b) GERD-PC, APP-NON-RES-% and GRANT-PMI.

contribute with less than or equal to 60% of the total applications for patents, while "middle" countries better exploit the help of foreign experts (see Figure 5.19a). The tree constructed from the GERD-PC and APP-RES-% presents semantically same relation, but with the opposite direction. The small subgroup of "low" countries within the deepest subtree in Figure 5.19a indicates that there exists a group of "low" countries, where more than 96.44% of patents are submitted by non-residents of the country (e.g., Democratic Republic of Congo, Sudan, Uganda, Madagascar, Sierra Leone). The core economic activities of these countries are agriculture and the extraction of natural resources. The development of their economies is largely supported by the IMF (International Monetary Fund), which may explain the help of foreign experts. Patents are most likely related to methods and tools for the extraction and processing of natural sources (e.g., petroleum, diamonds, gold, and copper) and to the improvement of the soil cultivation processes in order to increase the mass production of coffee, cocoa, tobacco, tea, etc.

Third, the number of granted patents emerged as a factor with an important influence on country's welfare. The tree constructed from GERD-PC, APP-NON-RES-% and "Grants of patents per million inhabitants" (GRANT-PMI) is of a quality higher than the initial tree (q_Δ of 0.1366). The tree contains credible relations (see Figure 5.19b), showing that more granted patents leads to better welfare. The relation, however, belongs to the second level of credibility, since it is not supported by the removed attributes graph (in Figure 5.17): each time the GRANT-PMI was removed, we did not observe a fall in quality.

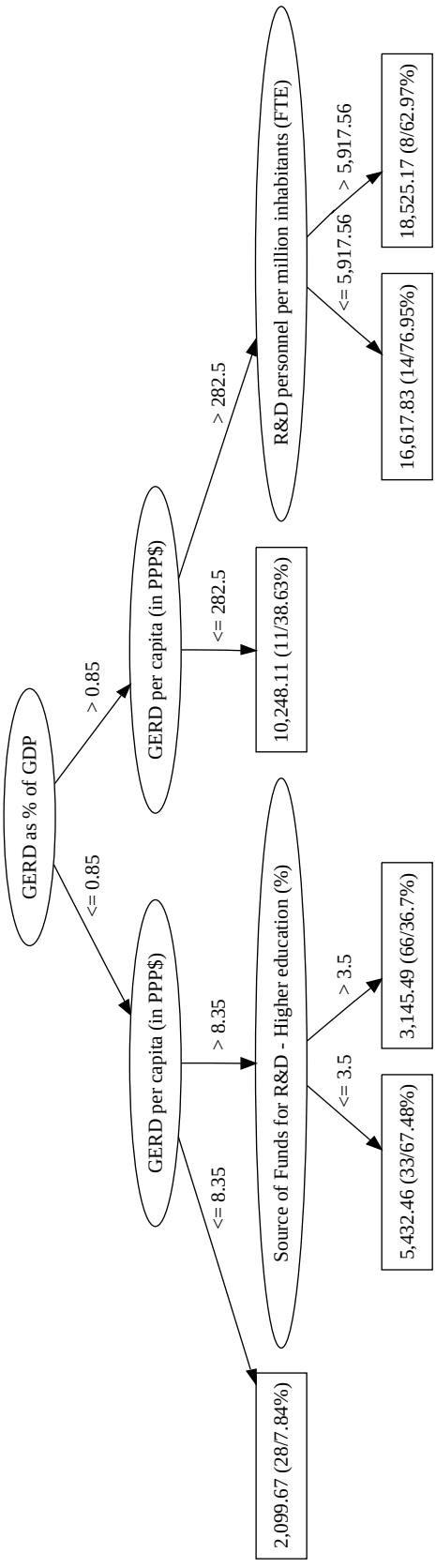
Fourth, the graph did not provide the support for those relations that describe the role of sectors: none of the trees containing one of the three SEC attributes contained credible relations. Therefore, the relations SEC-RES||SEC-R&D-PERS and SEC-INVEST belong to the second level of credibility. We observed several trees in order to define the relations between the attributes and a country's welfare. In general, to improve welfare, countries should increase the R&D activities within the business enterprise sector. For example, in "low" countries, government institutions have the leading role in R&D activities. It should be noted that government institutions do not include higher education institutions and research institutes, but those institutions that support common services to the community and administer the state and the economic and social policies. The trees constructed from the SEC-INVEST are of poor quality and cannot properly clarify the relations.

Step 7: Integrate conclusions.

The analyses made with the modified attributes set confirmed some of the existing and revealed some new relations. First, the GERD-PC||GERD-GDP relation was upgraded with the constructed attribute GERD-GNI and confirmed as a single first level of credibility relation. Second, further evidence was obtained, explaining the important role of patents. In other words, all countries should generally increase the number of patents (GERD-PC&APP-PMI and GRANT-PMI) to improve welfare. Third, it is not only important to increase the number of researchers (RES-HC||RES-FTE), but it is also important in which sector they are employed (SEC-RES||SEC-R&D-PERS). For example, while "low" countries have most of the researchers employed in the government, in "high" countries business enterprises employ the most researchers. Fourth, the relations HI-TECH and SEC-INVEST emerged at the third level of credibility.

5.2.3 Regression Trees Constructed from the Modified Attribute Set

Step 1: The modified attribute set from Subsection 5.2.2 is used, except that the discrete "GNI per capita" class is substituted with its numerical counterpart expressed in US$.



CC 0.722; RAA 33.5%; CPX 6

Figure 5.20: The initial tree constructed from the modified R&D attribute set with the regression trees.

Step 3,4: The same setup is used as in Subsection 5.1.3.
Step 5: The INITIAL_DM returned three trees, from which we selected the tree in Figure 5.20 constructed with the parameter MNIL 12.
Step 6: Modify the initial tree.

Remove attributes. The removed attributes graph presented in Figure 5.21 is constructed by first removing the attributes from the initial tree and then by removing a sequence of root attributes to the point at which there are no more interesting attributes. The graph supports already established relations and reveals two new findings. We will discuss only the new findings.

First, the graph provides an explanation for the SEC-INVEST relation. "Source of funds for R&D – Business enterprise (%)" (SRC-BE) attribute, which represents one of the sectors covered by the SEC-INVEST, emerges as the next best attribute when all three GERD attributes are removed. Within the tree, the SRC-BE characterizes "high" countries (those with the class value higher than or equal to 9,206 US$) as those countries where business enterprises provide more than 41.2% of the total investments in the R&D sector. When removed, the SRC-BE causes a further fall in quality (q_Δ of -0.2093) and its role is taken by the specific form of the SEC-INVEST attribute. Since M5P binarizes discrete attributes, the SEC-INVEST is transformed into the "SEC-INVEST = Business enterprise" form, indicating whether the business enterprise sector (> 0.5) or some other sector (≤ 0.5) invests the most in R&D. The attribute formed the same type of relation as the SRC-BE attribute, implying that it is important to stimulate business enterprises to increase the level of investment into R&D activities in order to improve welfare. The removal of SEC-INVEST causes a further fall in quality (q_Δ of -0.2412), which is sufficient evidence to establish the relation SEC-INVEST||SRC-BE.

Second, the graph reveals a new redundancy, semantically similar to the relation RES-HC||RES-FTE. The redundant relation is composed of two constructed attributes "R&D personnel per million inhabitants (FTE)" (R&D-PERS-FTE) and "R&D personnel per million inhabitants (HC)" (R&D-PERS-HC), which indicate the number of employees in the R&D sector per million inhabitants. When R&D-PERS-FTE is removed at the first level of the graph, the quality does not change because R&D-PERS-HC takes its role. The removal of both attributes causes a fall in all quality measures (q_Δ of -0.006), therefore indicating the redundant relation R&D-PERS-HC||R&D-PERS-FTE. The new relation is semantically similar to RES-HC||RES-FTE, since researchers generally form the majority of the R&D personnel.

Add attributes. The added attributes graph presented in Figure 5.22 is constructed with an expand credibility indicator tool by first adding all the attributes, and then by expanding the most promising indicators. We present only those parts of the graph, which provide new insights into the domain. The omitted parts confirm the relations discovered to this point, except the APP-NON-RES and HI-TECH relations, and are not further discussed. In total, there is one new finding: the graph provides an additional explanation for the R&D-PERS-HC||R&D-PERS-FTE relation. When any of the two attributes is added to the GERD attributes, the result is a tree of quality higher than the initial tree. However, the relation belongs to the second level of credibility, since it describes a subgroup. As can be seen from the tree in Figure 5.23, R&D-PERS-FTE differentiates between the two types of "high" countries (GNI per capita $\geq$ 9,206 US$), indicating that by increasing the number of people involved in R&D activities, the welfare can be even further improved. The same relation is obtained when R&D-PERS-FTE is substituted by R&D-PERS-HC.

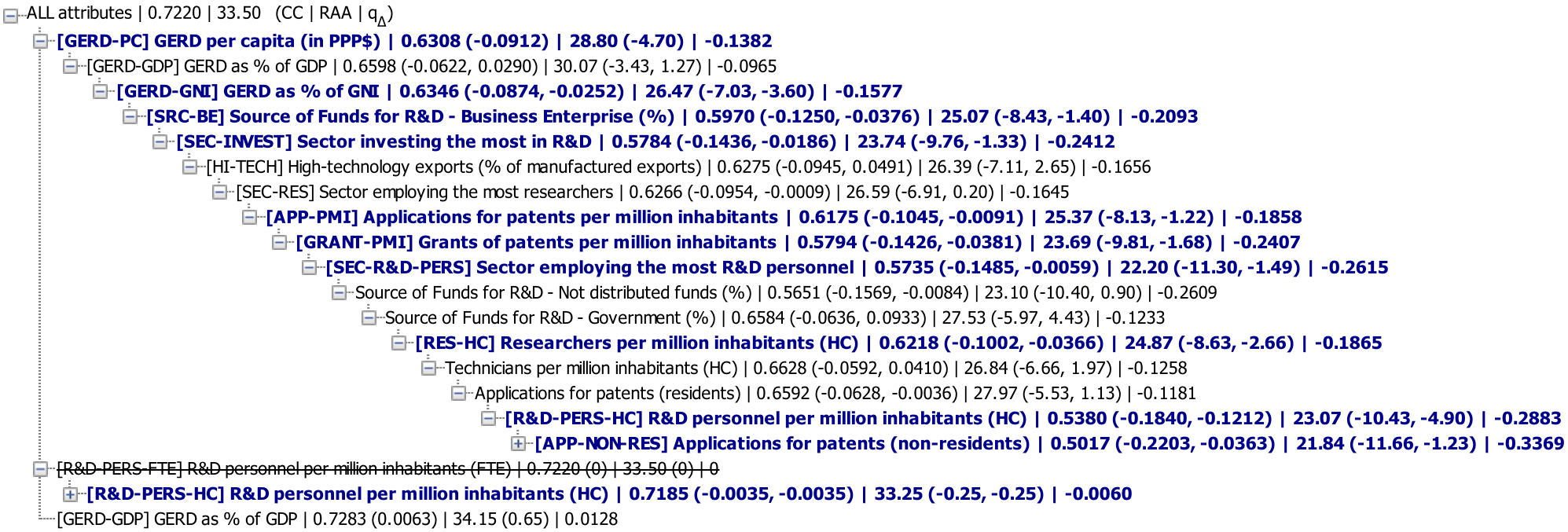


Figure 5.21: The removed attributes graph constructed from the modified R&D attribute set with the regression trees.

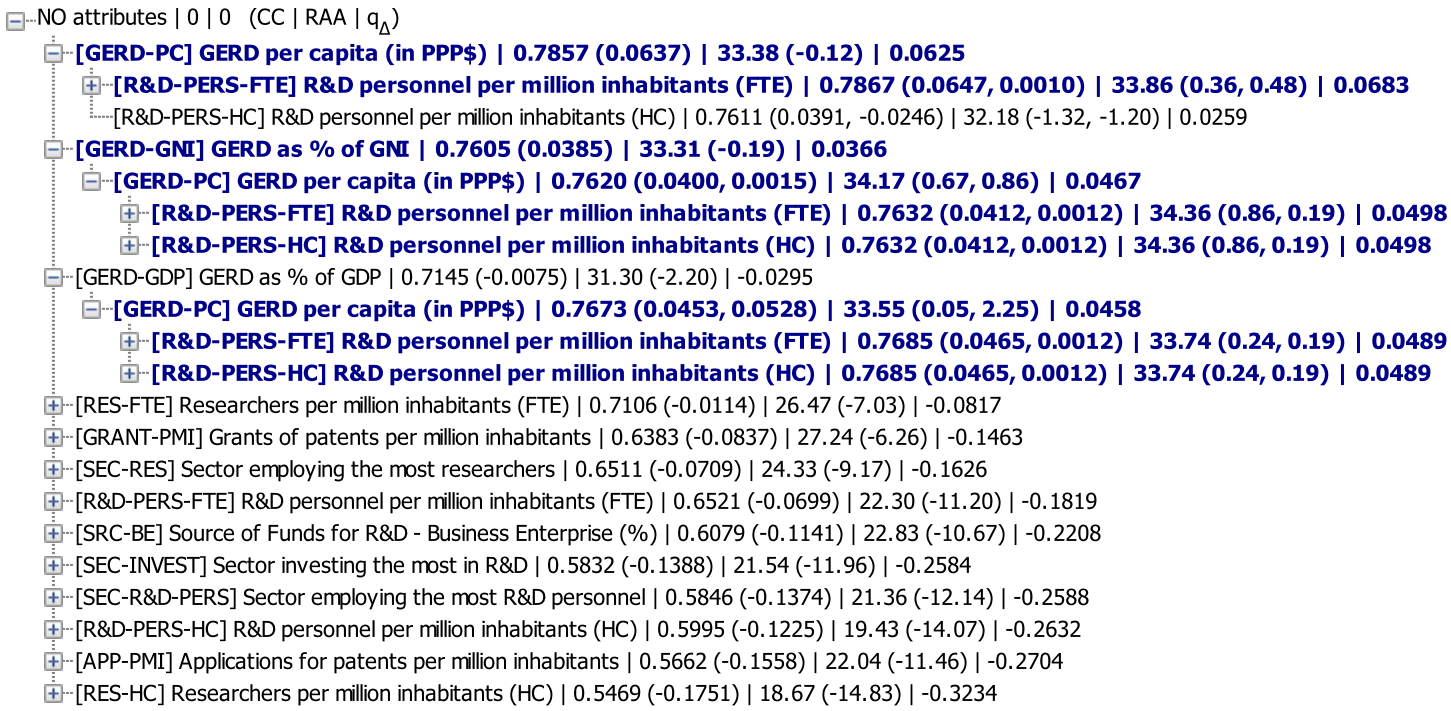


Figure 5.22: The added attributes graph constructed from the modified R&D attribute set with the regression trees.

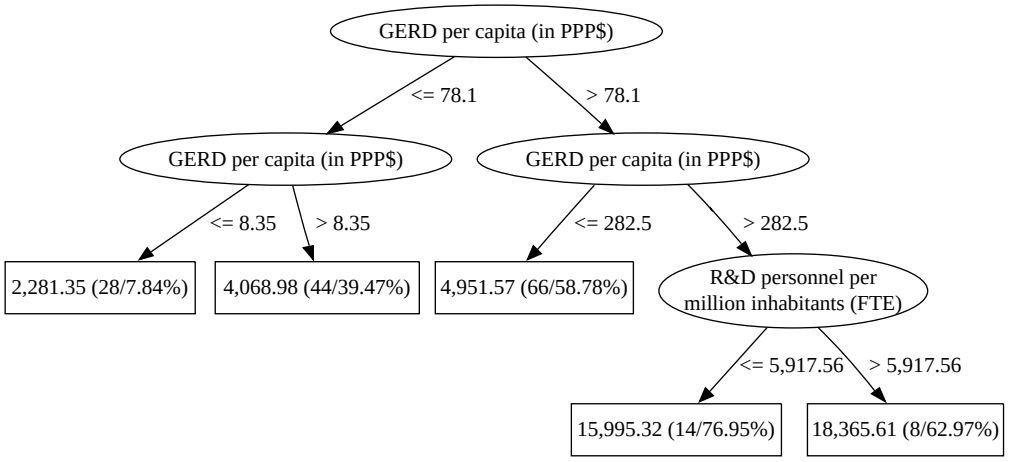


Figure 5.23: The credible tree constructed from GERD-PC and R&D-PERS-FTE.

Step 7: Integrate conclusions.

In order to improve welfare, the most important measure is to increase the level of investment in the R&D sector (GERD-PC||GERD-GNI||GERD-GDP). The following additional measures are required. First, the work on innovations should be intensified, which is reflected in a larger number of patents (APP-PMI, GRANT-PMI). Developing countries should better exploit the help of foreign experts in order to increase the number of patents (APP-NON-RES). Second, it is important to engage more people in R&D activities (RES-HC||RES-FTE, R&D-PERS-HC||R&D-PERS-FTE). Third, the business enterprise sector should be the key leader in R&D activities (SEC-RES||SEC-R&D-PERS, SEC-INVEST||SRC-BE). Finally, some evidence indicated that the amount of exported goods and services obtained as a result of intensive R&D activities (HI-TECH) is important for a country's welfare, but the direction of this relation remained unexplained. Some relations belonging to the third level appeared in the analysis, e.g., the number of female researchers. Analyses sometimes consider them as important; however, verifications revealed that their influence is weak.

The analyses with R&D attributes provide further proof that the HMDM method is capable of finding credible relations. The importance of the level of investment in R&D was acknowledged in the economic literature. It is generally used as a control variable (Varsakelis, 2006) to examine how successful the proposed method is at detecting credible attributes. The literature also supports the discovered relations of the second level of credibility that are presented in the previous paragraph (Furman et al., 2002).

In summary, this section presented the second application of the HMDM method. This time the HMDM was used to analyze the R&D data. The analysis showed that the most important measure to improve country's welfare is to increase the level of investment in R&D sector. The established relation is typically used in the literature to test the capabilities of the method to find important relations in the domain. An additional evaluation of the HMDM method is provided in the following section.

5.3 Evaluation of the Credible Models

Credible models should posses two additional properties. First, they should represent non-random patterns in data to support user's conclusions. Second, they should improve the initial model, not only in meaning, but also in quality.

Randomness was tested by comparing the credible models stored during the analysis with a random model. A model is considered as non-random when it is significantly better than the random model. In the case of decision trees, we consider as random the tree that always returns the majority class (in our case "middle") and in the case of regression trees, if it returns the mean of the actual values (in our case 1.93 for the higher education data and 6,258 for the R&D data).

Table 5.1: Comparison of credible decision trees with the baseline.

Decision trees	ACC	Diff.(PP)	CCPE	Kappa
Baseline	47.31%		0	0
GER-Total & GOER (Figure 5.4a)	74.85%	27.54	0.7047	0.5944
GER-Total & PE-GNI (Figure 5.4b)	73.65%	26.34	0.6767	0.5811
GER-Total + GOER & GOER & GER-Total + PE-GNI (Figure 5.5)	75.45%	28.14	0.7139	0.6060
GERD-PC (Figure 5.15a)	63.47%	16.16	0.5671	0.4023
GERD-GDP & APP-NON-RES (Figure 5.15b)	63.47%	16.16	0.5228	0.3993
GERD-PC & APP-NON-RES-% (Figure 5.19a)	70.06%	22.75	0.4741	0.5025
GERD-PC & APP-NON-RES-% & GRANT-PMI (Figure 5.19b)	70.66%	23.35	0.4733	0.5116

Table 5.2: Comparison of credible regression trees with the baseline.

Regression trees	CC	RAA
Baseline – Higher education data	−0.2775	0
GOER & GER (Figure 5.11a)	0.7291	24.55%
TERT-STUD & GER & OMR (Figure 5.11b)	0.6935	23.04%
Baseline – R&D data	−0.1274	0
GERD-PC & R&D-PERS-FTE (Figure 5.23)	0.7867	33.86%

Table 5.1 presents a comparison of the credible decision trees with the baseline. A direct comparison was made only for ACC, since for CCPE and Kappa any positive value indicates a non-random model. The increase in ACC (see the column "Diff. (PP)") varies from 16.16 to 28.14 percentage points (PP), which is significantly different from the baseline, strongly indicating that the effort in using our approach paid off. Positive values of the CCPE and Kappa provide further support for the conclusion.

Table 5.2 presents a comparison of the credible regression trees with the baseline. We did not straightforwardly compute the differences in CC since they are of different signs. Considering that "negative values should not occur for reasonable prediction models" (Witten and Frank, 2005), it is clear that our method was able to find non-random patterns in the data. Positive values of RAA provide an additional support to the conclusion.

The credible models are further compared to an initial model by measuring differences in quality between the first initial model constructed for a domain and each of the credible models constructed for the same domain. The first initial model is the closest approximation of the model obtained by the typical DM approach that constructs several models with different parameters and selects the best model. The comparisons for the decision

Table 5.3: Comparison of the credible decision trees with the initial tree.

Decision trees	ACC	Diff.(PP)	CCPE	Diff.	Kappa	Diff.
Initial model – Higher education data (Figure 5.1)	71.86%		0.5879		0.5497	
GER-Total & GOER (Figure 5.4a)	74.85%	2.99	0.7047	0.1168	0.5944	0.0447
GER-Total & PE-GNI (Figure 5.4b)	73.65%	1.79	0.6767	0.0888	0.5811	0.0314
GER-Total + GOER & GOER & GER-Total + PE-GNI (Figure 5.5)	75.45%	3.59	0.7139	0.1260	0.6060	0.0563
Initial model – R&D data (Figure 5.12)	63.47%		0.4438		0.3780	
GERD-PC (Figure 5.15a)	63.47%	0	0.5671	0.1233	0.4023	0.0243
GERD-GDP & APP-NON-RES (Figure 5.15b)	63.47%	0	0.5228	0.0790	0.3993	0.0213
GERD-PC & APP-NON-RES-% (Figure 5.19a)	70.06%	6.59	0.4741	0.0303	0.5025	0.1245
GERD-PC & APP-NON-RES-% & GRANT-PMI (Figure 5.19b)	70.66%	7.19	0.4733	0.0295	0.5116	0.1336

trees are presented in Table 5.3 and for the regression trees in Table 5.4. From the two tables, it can be seen that the credible models are generally of higher quality than the initial model. The conclusion is supported by the improvements in all quality measures: ACC (0-7.19 PP), CCPE (0.0295-0.1260), Kappa (0.0213-0.1336), CC (0.0128-0.0647) and RAA (0.36-3.16 PP).

Table 5.4: Comparison of the credible regression trees with the initial tree.

Regression trees	CC	Diff.	RAA	Diff.(PP)
Initial model – Higher education data (Figure 5.8)	0.6807		21.39%	
GOER & GER (Figure 5.11a)	0.7291	0.0484	24.55%	3.16
TERT-STUD & GER & OMR (Figure 5.11b)	0.6935	0.0128	23.04%	1.65
Initial model – R&D data (Figure 5.20)	0.7220		33.5%	
GERD-PC & R&D-PERS-FTE (Figure 5.23)	0.7867	0.0647	33.86%	0.36

In summary, this chapter presented two applications of the HMDM method, showing how higher education and R&D sectors influence the economic welfare of a country. The credible relations obtained by the HMDM were compared to the relations established in the literature. The results of comparison showed that our method is capable of finding important relations in a domain. Finally, we showed that the credible models discovered by the HMDM represent non-random patterns in data, which improve the models constructed with the typical DM approach.

6 Learning Predictive Models with the HMDM

In contrast to the previous chapter, where the emphasis was on finding credible relations, in this chapter we present an application of the HMDM for the construction of credible predictive models. In this light, the user performs the HMDM in order to obtain a set of credible decision tree models, which are combined in an ensemble and used for the prediction of previously unseen instances. The predictive capabilities of the HMDM are demonstrated in the domain of automatic web genre identification (AWGI) and compared with the state-of-the-art ML algorithm for the construction of decision trees.

The chapter is organized in four sections. Section 6.1 presents the problem of AWGI and the data. Section 6.2 describes the HMDM upgrade, made to construct a multi-label classifier from the AWGI data. Experimental design is presented in Section 6.3, while the chapter concludes with results and discussion (Section 6.4).

6.1 Automatic Web Genre Identification

6.1.1 The Task of AWGI

Genre is a complex concept, used in literary studies, linguistics and rhetoric. Furthermore, it is applicable even to non-textual media, such as movies and music. In spite of the prevalence in many disciplines, there is no overall consensus of what genre is (Santini, 2007). In the field of AWGI, genre is generally considered as a specific typology of web documents, which defines the style of a web page (Karlgren, 2000). For example, "Blog" is a genre that presents updates on what is going on with an entity. The "Blog" page typically presents the updates as series of articles named posts, followed by the comments of people who follow the updates. This style is independent of the topic presented on the "Blog" page.

AWGI is useful in information retrieval systems and digital libraries. For example, by specifying genres beside keywords in information retrieval systems, the quality of the retrieved results can be significantly improved (Vidulin et al., 2007d).

One of the key research topics in AWGI is an analysis of genre-specific attributes in order to find an appropriate set of automatically extractable attributes. For example, Shepherd et al. (2004) defined web genres with a triplet <content, form, functionality> representing three dimensions of genres, which serve as a framework for the choice of attributes: a) typical content attributes would be, e.g., content words, function words and punctuation signs; b) the form is reflected in attributes describing a syntactic form of text and the design of a web page; c) the functionality addresses the interactive capabilities offered by web pages, such as hyperlinks.

In our previous work (Vidulin et al., 2006, 2007b,c,a,d; Luštrek et al., 2007; Vidulin et al., 2009; Vidulin, 2009; Vidulin and Gams, 2009) we experimented with all three types of attributes suggested by Shepherd et al. (2004) and observed that content words may have an important role in recognizing web genres. Therefore, the aim of the experiments presented in this chapter is to explore and define the role of content words in the AWGI.

6.1.2 20-Genre Corpus

In order to construct predictive models from data, we used the 20-Genre corpus (available at http://dis.ijs.si/mitjal/genre/). The corpus consists of 1,539 web pages in English classified into 20 genres. Considering that the corpus was gathered from the internet, where genres are far from clearly delineated, the corpus is multi-labelled. In other words, each web page can belong to multiple genres. In total, from the 1,539 web pages, 1,059 are labelled with one, 438 with two, 39 with three and 3 with four labels. On average, there are 1.34 labels per web page.

Table 6.1: The composition of 20-Genre corpus.

Genre	**Web pages**
Blog	77
Children's	105
Commercial/promotional	121
Community	82
Content delivery	138
Entertainment	76
Error message	79
FAQ	70
Gateway	77
Index	227
Informative	225
Journalistic	186
Official	55
Personal	113
Poetry	72
Pornographic	68
Prose fiction	67
Scientific	76
Shopping	66
User input	84

The composition of the corpus is presented in Table 6.1. "Blog" presents updates on what is going on with an entity. "Children's" presents content in a simple and colourful way specifically suited for children. "Commercial/promotional" pages are intended to invoke the visitor's interest in goods or services, typically for commercial gain. "Community" type page involves the visitor in the creation of content and enables interaction with other visitors. "Content delivery" delivers content that is not a part of the page. "Entertainment" pages entertain the visitor. "Error message" tells the visitor to go away. "FAQ" are intended to help a user to solve common problems by answering frequently asked questions. "Gateway" transfers the visitor to another page. "Index" transfers the visitor to a selection of multiple other pages. "Informative" conveys objective information of permanent interest suitable for the general population. "Journalistic" conveys mostly objective information on current events. "Official" conveys information with legal or otherwise official consequences. "Personal" conveys subjective, personal information in an informal way. "Poetry" presents poems and lyrics with the intention to evoke emotions. "Pornographic" web pages have the intention to sexually arouse the visitor. "Prose fiction" presents a story about a real or fictional event in artistic form with the intention to evoke imagination and emotions. "Scientific" conveys objective information suitable

for experts. "Shopping" web pages sell goods or services online. "User input" solicits the visitor's input.

The presented genre categories form a coarse-grained scheme, intended to cover diverse pages available on the internet. Accordingly, categories as "Error message" are defined in order to enable the user to filter out useless pages.

6.1.3 Data Preparation

A data set is obtained from the 20-Genre corpus in five steps. First, a text is extracted from each web page. Second, a set of words is extracted from the texts, where a word is treated as a sequence of letters. Third, stop-words are removed. We used the list of stop-words presented in (Lewis et al., 2004). Fourth, the data set is composed, where attributes are the extracted words, instances are the web pages from the corpus and values represent the presence or absence of the words in the pages. This type of representation is named bag-of-words and is commonly used for representing textual documents (Sebastiani, 2002). An alternative representation would be tfidf (Salton and Buckley, 1988). However, we decided to use the bag-of-words to simplify the task for the user executing the HMDM. In this manner, the user deals only with the presence or absence of words and not with the degrees of presence. Finally, the class is attached, which is represented with an attribute vector $\mathbf{Y}$ composed of binary attributes (20 in our case) that indicate genres to which the web page belongs.

The data set is divided into half using a stratified split. The first part is intended for training and the second for testing.

6.2 Construction of a Multi-Label Classifier with the HMDM

The task of AWGI that we demonstrate is a multi-label classification task, which associates an instance $\mathbf{X}$ with a subset of labels $\mathbf{Y} \subseteq L$, where L represents a previously known finite set of labels.

To construct a multi-label classifier with the HMDM, we applied the *binary relevance problem transformation method* (BRPT) (Tsoumakas et al., 2010) presented in Figure 6.1. The BRPT transforms a multi-label data set into $|L|$ binary data sets by: first, copying all attribute-value pairs $|L|$ times; then, transforming labels of examples (EX#) into positive (true) if they belong to the class of interest, and into negative (false) if they belong to any other classes. For each binary data set, a binary sub-classifier is constructed with an arbitrary single-label ML algorithm, which is in our case the HMDM. Finally, the sub-classifiers are added to a multi-label classifier. When the multi-label classifier classifies an instance, the instance is classified with all sub-classifiers and the resulting classification is a subset of classes for which the sub-classifiers returned a positive answer.

The procedure that combines the HMDM with the BRPT, further referred to as HMDM_{ML}, is presented in Algorithm 6.1. In short, for each class (genre in our case) a binary data set D_i is extracted from the multi-label data set D. Then, the attribute selection method is applied to obtain a manageable subset of attributes, as common in text categorization tasks. For this purpose, the attributes are ranked in decreasing order of information gain and the top 500 attributes from over 20,000 are selected. Finally, a sub-classifier is constructed with the HMDM and included into the multi-label classifier. When the output of the HMDM contains multiple models, they are fused into an ensemble by a majority vote function.

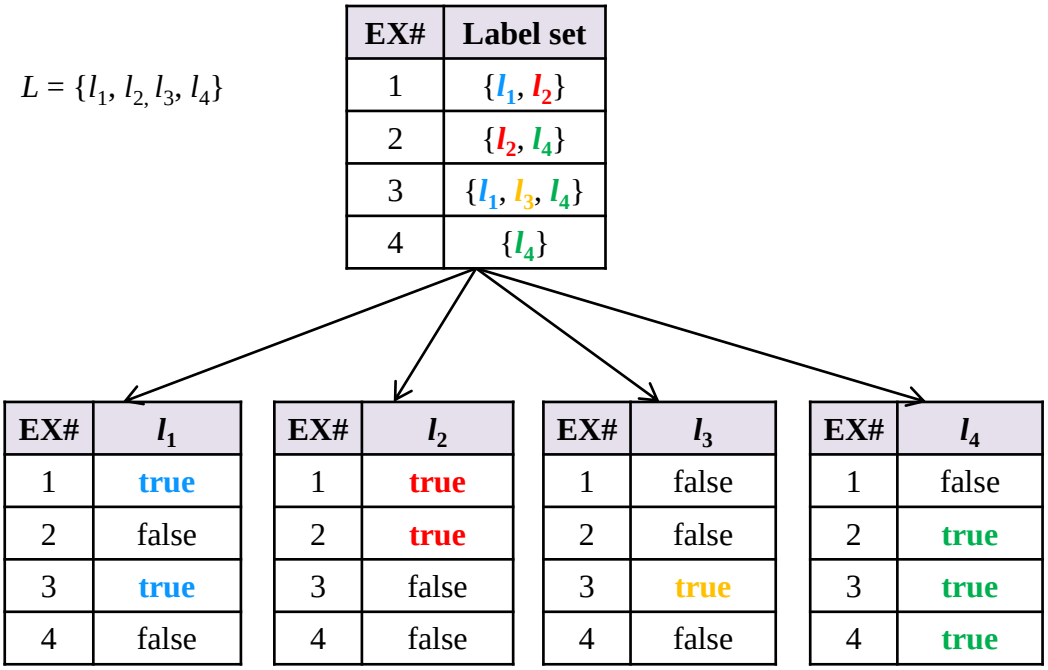


Figure 6.1: Binary relevance problem transformation method.

Algorithm 6.1: The HMDM_{ML} – a multi-label variant of the HMDM algorithm.

HMDM_{ML} (a multi-label data set D)
1 FOR each distinct class l_i from D
2 Extract from D a binary-class data set D_i, which represents the class l_i
3 Rank attributes in D_i in decreasing order of information gain and select the top 500 attributes
4 Perform the HMDM on D_i in order to obtain a sub-classifier C_i
5 Add C_i to a multi-label classifier C_{ML}
6 END FOR
8 Return the multi-label classifier C_{ML}

CLASSIFY (an instance $\mathbf{X}$, the multi-label classifier C_{ML})
1 FOR each distinct class l_i from L
2 Classify $\mathbf{X}$ with C_i in order to obtain the classification $\mathbf{Y_i}$
3 END FOR
4 Return $\mathbf{Y}$

The second procedure in Algorithm 6.1 explains how a new instance is classified with the multi-label classifier. In short, the instance is classified by $|L|$ sub-classifiers and the resulting classification is a subset of classes for which the sub-classifiers returned a positive answer.

6.3 Experimental Design

The experiments are designed to answer the following research questions: Q1) Assuming that credible models constructed by the HMDM contain only meaningful and high-quality[1] relations, will they also exhibit high predictive performance on previously unseen instances?; Q2) Are web genres, represented within 20-Genre corpus, describable with content words and if they are, which genre-specific words are the most influential?

To answer Q1, the predictive performance of credible decision trees, measured on a separate test set, is compared against the predictive performance of the J48 decision trees, measured on the same set. Considering that the HMDM is aimed at improving the J48 decision trees, the credible trees should exhibit higher predictive performance on the separate test set.

To answer Q2, a formal analysis of 20 genre categories was conducted in order to divide genres into those describable with content words and those for which other types of attributes should be used. The former group contains those genres for which meaningful trees with an acceptable predictive performance can be constructed. In contrast, the latter group contains those genres for which either it was not possible to construct meaningful trees, or the meaningful trees exhibit poor predictive performance.

To construct credible trees, the HMDM was conducted with the same set of parameters used in the experiments presented in Chapter 5. The J48 tree was constructed with the default parameters set in Weka.

The predictive performance of the trees is estimated both at the level of binary sub-classifiers and at the level of a multi-label classifier. The choice of measures was restricted by the specific properties of the data transformed by the BRPT. In other words, the resulting binary data sets exhibit a significant disbalance between the number of positive and negative examples, where negative examples dominate. In such a setting, ACC is an inappropriate measure, since models able to recognize only negative examples would still have a high ACC. Therefore, we used tree measures common to text categorization tasks, which do not account for negative examples: precision, recall and F-Measure (Sebastiani, 2002). In the context of AWGI, precision can be interpreted as a ratio of web pages a sub-classifier correctly classified to a genre it represents and all pages the sub-classifier attributed to the same genre. Similarly, recall can be interpreted as a ratio of pages correctly classified to a genre and all the pages that actually belong to the genre. F-Measure represents a combination of the two measures. The three measures are distributed within the $[0, 1]$ interval, with a higher number denoting a better sub-classifier.

Formally, F-Measure is defined as:

$$F_1 = 2\frac{\pi\rho}{\pi + \rho}, \tag{6.1}$$

and represents a harmonic mean of precision (Eq. 6.2) and recall (Eq. 6.3):

$$\pi = \frac{TP}{TP + FP}, \tag{6.2}$$

[1]Quality was assessed with the help of 10-fold cross-validation on a train set.

$$\rho = \frac{TP}{TP + FN}, \tag{6.3}$$

where TP stands for *true positives* and represents a number of web pages belonging to the class $l_i \in L$ that are classified as such. FP stands for *false positives* and represents a number of non-l_i pages classified as l_i. Finally, FN stands for *false negatives* and represents a number of l_i pages classified as non-l_i.

The predictive performance of the multi-label classifier is also represented by precision, recall and F-Measure, however, averaged over all the classifier's decisions (micro-averaging) and over all the categories (macro-averaging).

Micro-averaged measures weigh all the web pages equally, representing the averages over all the (web page, genre category) pairs. They tend to be dominated by the classifier's performance on common categories. The micro-averaged precision π(micro) represents the ratio of web pages correctly classified as l (TP_l = true positives), and all the pages correctly and incorrectly (FP_l = false positives) classified as l (Eq. 6.4). Micro-averaged recall ρ(micro) represents the ratio of pages correctly classified as l, and all the pages actually pertaining to the class l (FN_l = false negatives) (Eq. 6.5). Micro-averaged F-measure F_1(micro) represents a harmonic mean of π(micro) and ρ(micro) (Eq. 6.6).

$$\pi\,(\text{micro}) = \frac{\sum_{l=1}^{|L|} TP_l}{\sum_{l=1}^{|L|}(TP_l + FP_l)}. \tag{6.4}$$

$$\rho\,(\text{micro}) = \frac{\sum_{l=1}^{|L|} TP_l}{\sum_{l=1}^{|L|}(TP_l + FN_l)}. \tag{6.5}$$

$$F_1\,(\text{micro}) = \frac{2 \times \pi\,(\text{micro}) \times \rho\,(\text{micro})}{\pi\,(\text{micro}) + \rho\,(\text{micro})}. \tag{6.6}$$

.

Macro-averaged measures weigh equally all the genre categories, regardless of their frequencies. They tend to be dominated by the classifier's performance on rare categories. Macro-averaged precision π(macro) is computed first by computing the precision for each category separately, and then by averaging over all the categories (Eq. 6.7). The same procedure is used for computing the macro-averaged recall ρ(macro) (Eq. 6.8), and macro-averaged F-measure F_1(macro) (Eq. 6.9).

$$\pi_l = \frac{TP_l}{TP_l + FP_l}, \quad \pi\,(\text{macro}) = \frac{\sum_{l=1}^{|L|} \pi_l}{|L|}. \tag{6.7}$$

$$\rho_l = \frac{TP_l}{TP_l + FN_l}, \quad \rho\,(\text{macro}) = \frac{\sum_{l=1}^{|L|} \rho_l}{|L|}. \tag{6.8}$$

$$F_{1l} = \frac{2 \times \pi_l \times \rho_l}{\pi_l + \rho_l}, \quad F_1\,(\text{macro}) = \frac{\sum_{l=1}^{|L|} F_l}{|L|}. \tag{6.9}$$

The credible trees are further compared with the J48 trees using McNemar's test (Dietterich, 1998). This test is suitable for making predictive-performance-based comparisons on the level of sub-classifiers. Let state that $\widehat{f_{HMDM_i}}$ is an ensemble constructed by the HMDM for i-th genre and $\widehat{f_{J48_i}}$ the tree constructed by the J48. For each genre, the

classifications made with both $\widehat{f_{HMDM_i}}$ and $\widehat{f_{J48_i}}$ are recorded and the following values are computed:

- n_{10_i} – the number of examples correctly classified by the $\widehat{f_{HMDM_i}}$ and incorrectly classified by the $\widehat{f_{J48_i}}$,
- n_{01_i} – the number of examples incorrectly classified by the $\widehat{f_{HMDM_i}}$ and correctly classified by the $\widehat{f_{J48_i}}$.

For the i-th sub-classifier, the test statistic, distributed as χ^2 with one degree of freedom, is computed using the following equation:

$$\chi^2 = \frac{(|n_{10_i} - n_{01_i}| - 1)^2}{n_{10_i} + n_{01_i}}. \tag{6.10}$$

Two adjustments of the presented test are applicable to multi-label classifiers. First, the test statistic for a multi-label classifier can be computed using macro-averaging, in the same way as for the quality measures. In other words, the test statistic is computed separately for each genre and averaged over all genres. Second, similar to micro-averaging, the n_{10_i} and n_{01_i} can be summed over all genres resulting in n_{10} and n_{01}, which are then included in the Eq. 6.10. We used both tests.

6.4 Results and Discussion

The results of predictive-performance-based comparisons (see Table 6.2) show that the credible trees exhibit a significantly better predictive performance than the J48 trees – the probability of making an error by stating that the HMDM performed better than the J48 on the AWGI task is less than or equal to 0.0124. This finding provides a positive answer to the Q1. Considering F-Measure, the credible trees performed better than the J48 trees for 15 genres, worse for four genres and equally well for one genre. The analysis of precision and recall reveals that the improvement is mostly due to the increase in precision of the credible trees. The same conclusion is supported by the micro- and macro-averaged measures (see the bottom rows of the table).

6.4.1 Explanation

A further analysis provided an explanation for the differences in performance achieved by the credible and J48 trees, and gave an answer to Q2. The variations in performance of credible trees are best explained through illustrative examples. The first example presents the analysis of genre "Blog", where credible trees are composed of more meaningful relations of considerably higher predictive performance (F-Measure of 0.738) than the J48 tree (0.406). In contrast, the second example presents the analysis of genre "Index", where the attempts to improve the J48 tree in meaning did not produce a tree that afterwards exhibited an acceptable predictive performance. This is typical for genres where content words are not suitable attributes.

Blog

A meaningful tree, which properly represents "Blog", should contain relations that differentiate a "Blog" page of any topic from pages of other genres. In other words, the tree should contain such relations, which are composed of genre-specific words.

Table 6.2: The results of predictive-performance-based comparisons between the decision trees constructed with the HMDM and J48.

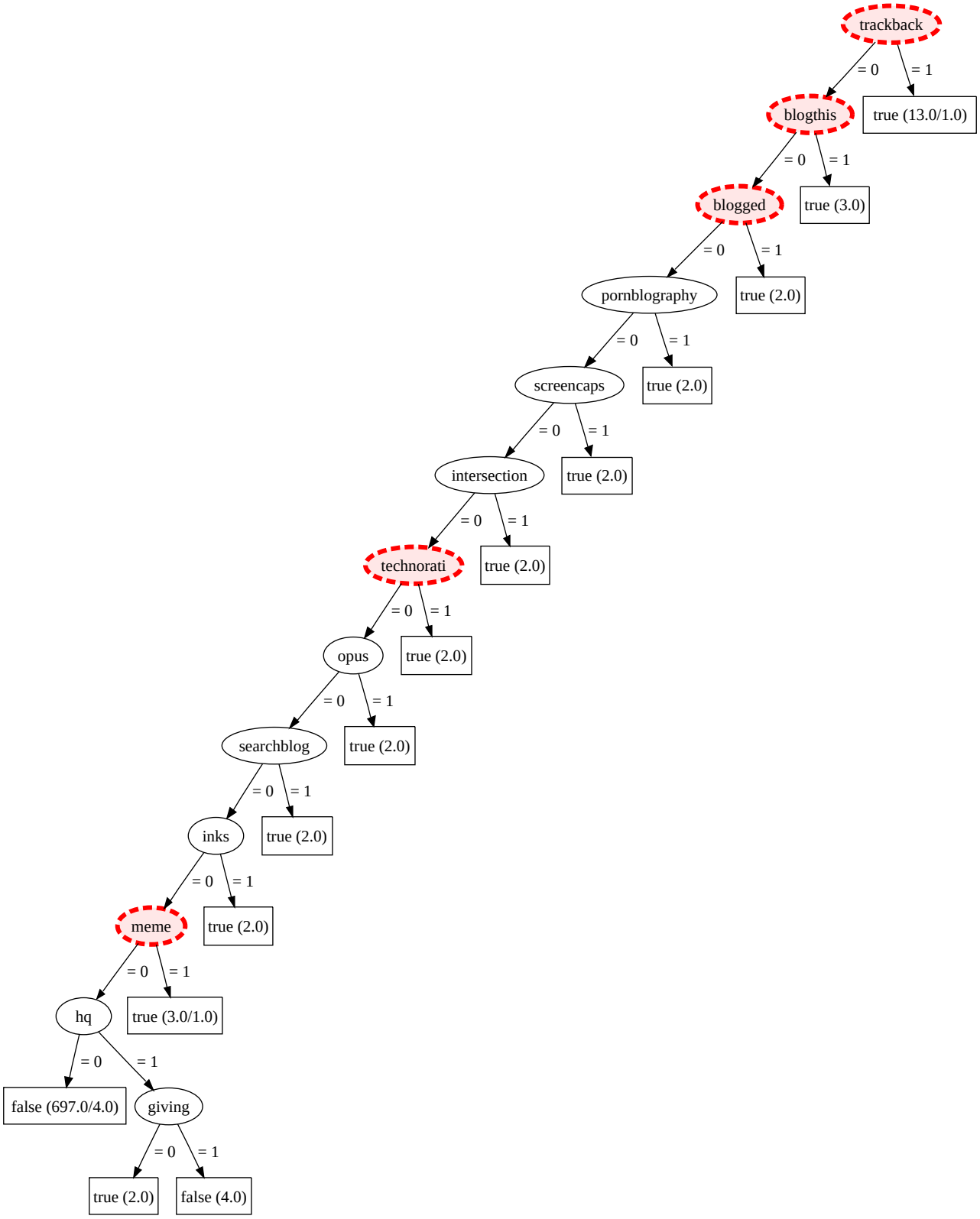
	J48 trees			**HMDM trees**		
GENRE	π	ρ	F_1	π	ρ	F_1
Blog	0.452	0.368	0.406	0.889	0.632	0.738
FAQ	0.438	0.206	0.280	0.633	0.559	0.594
Community	0.472	0.415	0.442	0.600	0.439	0.507
Error message	0.541	0.556	0.548	0.760	0.528	0.623
Shopping	0.381	0.242	0.296	0.636	0.212	0.318
Scientific	0.320	0.211	0.254	0.533	0.421	0.471
Poetry	0.395	0.417	0.405	0.459	0.472	0.466
Official	0.500	0.296	0.372	0.600	0.333	0.429
Pornographic	0.560	0.412	0.475	0.692	0.794	0.740
Commercial-promotional	0.200	0.034	0.058	0	0	0
Index	0.365	0.176	0.238	0.667	0.037	0.070
Prose fiction	0.472	0.515	0.493	0.667	0.606	0.635
Children's	0.857	0.261	0.400	0.458	0.587	0.514
Journalistic	0.450	0.290	0.353	0.604	0.344	0.438
Gateway	0	0	0	0	0	0
Entertainment	0.065	0.054	0.059	0.692	0.243	0.360
Personal	0.417	0.089	0.147	0.283	0.232	0.255
Content delivery	0.059	0.016	0.026	0	0	0
Informative	0.237	0.126	0.165	0.500	0.045	0.083
User input	0.583	0.171	0.264	1	0.171	0.292
	(wins / losses / ties):			15/4/1	12/6/2	15/4/1
Micro-AVG	0.390	0.219	0.280	0.564	0.275	0.370
Macro-AVG	0.388	0.243	0.284	0.534	0.333	0.377
McNemar's – Micro-AVG	$\chi^2 = 58.343$, $p < 0.0001$					
McNemar's – Macro-AVG	$\chi^2 = 6.243$, $p = 0.0124$					

The J48 tree is presented in Figure 6.2. A node in the tree represents a word, the left branch (= 0) represents the absence and the right branch (=1) the presence of the word in the web page. Leaves containing "true" indicate a "Blog" web page and leaves containing "false" indicate a non-"Blog" web page. The measures below the tree denote the tree's quality estimated by 10-fold-cross validation executed while performing the HMDM.

The J48 tree contains 14 relations, which state that when one of the 12 words – trackback, blogthis, blogged, pornblography, screencaps, intersection, technorati, opus, searchblog, inks, meme, hq – appears in the web page, the web page is an instance of "Blog". In contrast, the absence of the word "giving" also indicates the "Blog" page.

The analysis shows that the J48 tree in Figure 6.2 only partially contains genre-specific words (the emphasized nodes). Considering the meaning of words and the context in which the words appear within the "Blog" web pages, the following words can be denoted as genre-specific:

- trackback – Represents a linkback method for blog authors to request a notification when somebody links to one of their pages. For example, the word "trackback" appears on the "Blog" page in the context of a counter, which indicates how many



F_1 0.4667; CCPE 0.8637; Kappa 0.4462; CPX 14

Figure 6.2: The J48 tree constructed for the genre “Blog”. The emphasized nodes denote genre-specific words.

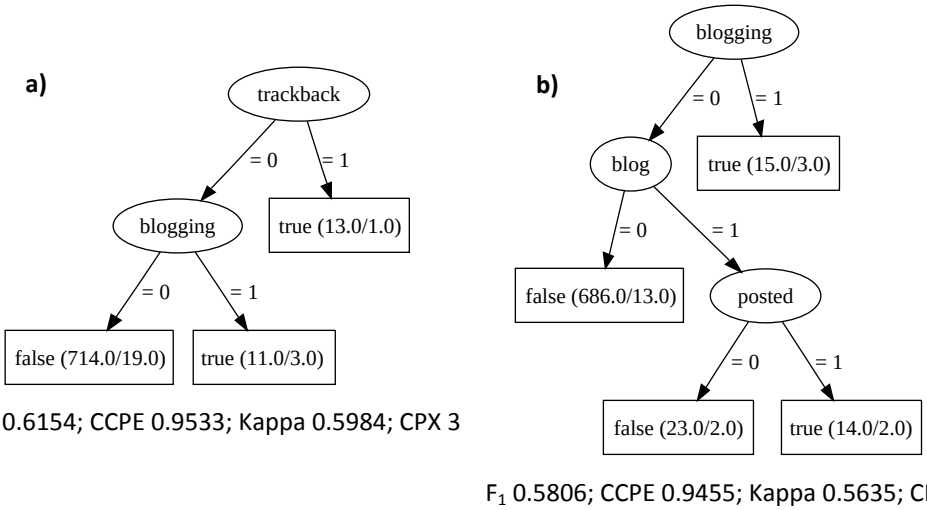
documents linked back to a specific blog post (e.g., "trackback (5)"). Another example is a direct offer to trackback, such as "trackback link is ...".

- blogthis – A service of blogger.com for writing a blog post without visiting the blogging site. It appears on the "Blog" pages mostly within heading, offering a direct access to the service.
- blogged – An act of adding new material to a blog. For example, "Yesterday, I blogged about ..."
- technorati – An internet search engine for user-generated media including blogs. For example, it appears as a counter indicating how many visitors found the specific blog using technorati.
- meme – Represents toughs propagated from one blog to another, often directly referred as memes.

Other words (pornblography, screencaps, intersection, opus, searchblog, inks and hq) denote specific blogs and are not general representatives of "Blog" pages.

In total, from 12 words denoted by the J48 tree as blog-specific, only five can be characterized as such, indicating that the tree can be further improved in terms of meaning.

By applying the HMDM, two trees are obtained. The first, presented in Figure 6.3a, contains relations stating that the "Blog" page should contain either one of the two words: trackback or blogging. The second, presented in Figure 6.3b characterizes the "Blog" page as the page containing either the word "blogging" or the combination of words "blog" and "posted".



F_1 0.6154; CCPE 0.9533; Kappa 0.5984; CPX 3

F_1 0.5806; CCPE 0.9455; Kappa 0.5635; CPX 4

Figure 6.3: The credible trees constructed with the HMDM for the genre "Blog".

All three quality measures presented below the credible trees in Figure 6.3 indicate that the two trees are of a quality higher than the J48 tree (see measures below the tree in Figure 6.2 for comparison). At the same time, the credible trees are less complex (CPX of 3 and 4 in comparison to the CPX of 14 in the case of the J48 tree).

The credible trees have only one word in common with the J48 tree: trackback. The other words contained within the credible trees are:

- blogging – It has the same meaning as "blogged", since both words represent the verb "to blog", which is once stated as a present participle and in other case as a past participle. It seems, however, that the form "blogging" is much more common. For example, in contexts such as "I've not been blogging much this week.".

- blog – The authors often directly refer to their blogs. For example, "Welcome to my blog!" or "The blog of seldom seen photography".
- posted – Denotes an act of displaying a post on the blog. For example, "This entry was posted on Sep 7, 2011" or "posted by ...".

In conclusion, the advantages of the credible trees are twofold. First, they are composed of concise and meaningful relations, providing straightforward explanations of blog-specific words. Second, by eliminating less-credible relations, a considerable increase in predictive performance is achieved (F-Measure of 0.738 in comparison to the F-Measure of 0.406 achieved by the J48 tree).

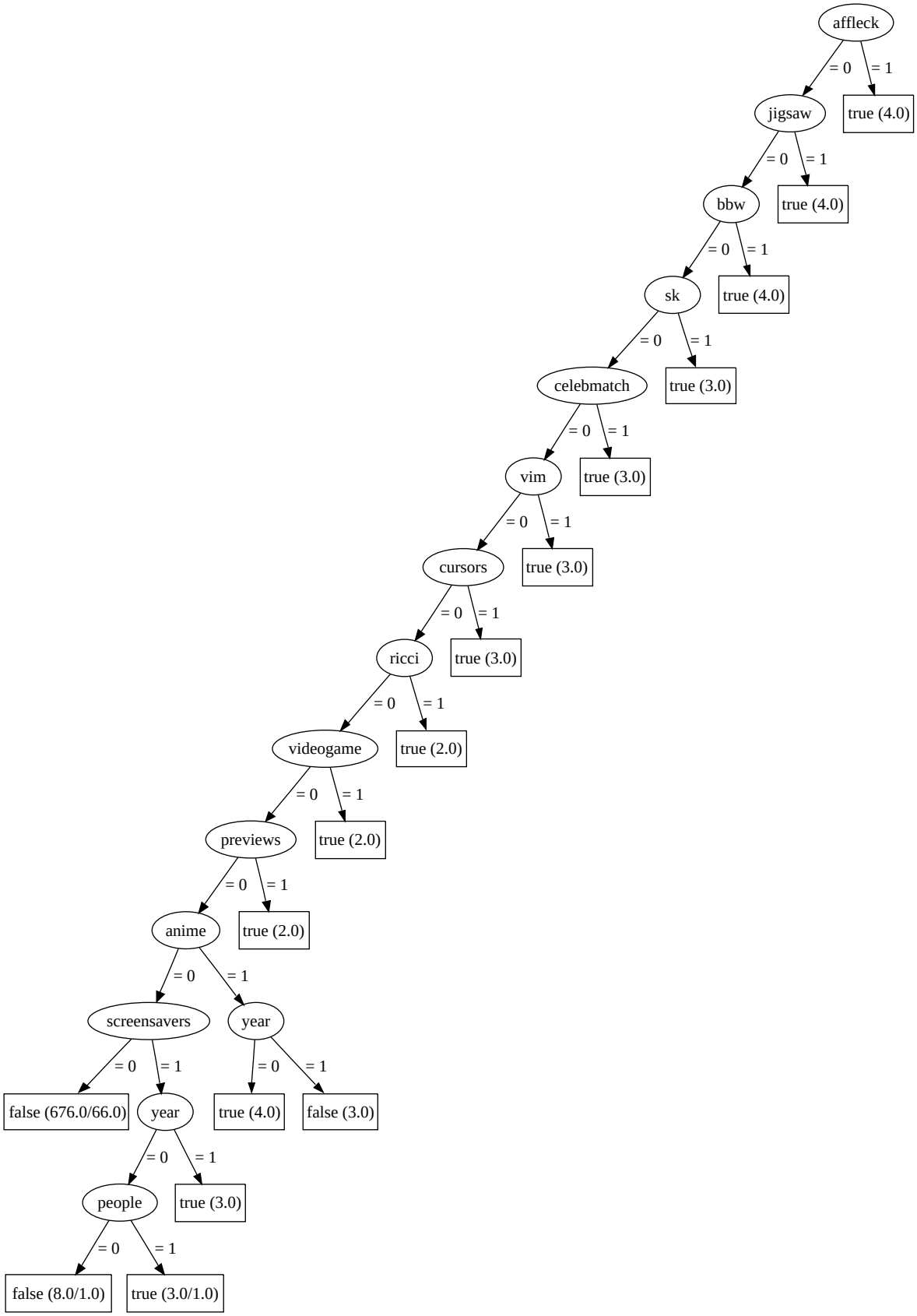
Index

An "Index" web page contains a collection of links or a table of content. One can expect that such a category can be described with words that introduce a list of links, such as "links" or "hotlist". In the case of a table of contents, the words "table" and "contents" could appear to introduce the purpose of the page.

The analysis of the J48 tree in Figure 6.4 indicates that although it exhibits a higher predictive performance (F-Measure of 0.238) than the credible tree (0.07), it contains meaningless relations, composed of words denoting topics of pages behind the links. This can be seen from the explanations of words that appear within the J48 tree:

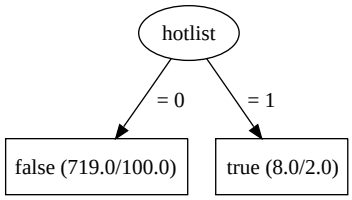
- affleck – Surname of an American actor Ben Affleck.
- jigsaw – A puzzle.
- bbw – An abbreviation of big beautiful woman, which appears within the indexes of pornographic pages.
- sk – Obtained by extracting a sequence of letters from "sk8t' Gamer Wallpaper".
- celebmatch – The link to an entertaining site, where the user can find the most compatible celebrity love based on his/her date of birth.
- vim – Appears in the context "designed with vim", where "vim" represents a link to vim editor.
- ricci – Surname of an American actress Christina Ricci.
- cursors, videogame, previews, anime, screensavers, year, people – Different topics.

With the help of the HMDM, we eliminated the topic-specific words and obtained a meaningful tree presented in Figure 6.5. The tree was only partially supported with quality measures within the HMDM: with higher CCPE (0.8521 in comparison to the CCPE of the J48 tree of 0.5117) and lower complexity (2 in comparison to the J48 tree's CPX of 16). However, it was the best observed tree along the two dimensions of meaning and quality. Considering the predictive performance, the precision of the credible tree (0.667) is considerably better than the precision of the J48 tree (0.365), while the recall is considerably lower (0.037 for the credible and 0.176 for the J48 tree). Higher precision indicates that "hotlist" represents a good choice of index-specific word, since between the pages containing the word "hotlist" there is a small number of pages that do not belong to the "Index" genre. In contrast, low recall indicates that all the pages belonging to this category cannot be properly recognized based on a single word.



F_1 0.3022; CCPE 0.5117; Kappa 0.2503; CPX 16

Figure 6.4: The J48 tree constructed for the genre "Index".



F_1 0.1053; CCPE 0.8521; Kappa 0.0866; CPX 2

Figure 6.5: The credible tree constructed with the HMDM for the genre "Index".

The presented example is a typical representative of genres not describable with words. Considering the lack of genre-specific words within the train set, the meaningful trees either cannot be found (F-Measure equals 0) or are of poor predictive performance, such as the tree in Figure 6.5. In comparison to the J48 tree, such an output of the HMDM has a lower predictive performance. Although one can expect that even the J48 tree that contains less-credible relations, which cannot be further improved, would exhibit a low predictive performance on another set of instances independent of the train and test sets.

Discussion

The two illustrative examples represent two types of credible models. The first type contains those credible models that achieve a better predictive performance than the models constructed with automatic methods, due to a successful elimination of the less-credible relations. The second type contains credible models that cannot achieve an acceptable predictive performance because of the lack of credible attributes in the data. The analysis of web pages representing the second group of genres indicated that they would be better described with attributes representing form and functionality. For example, a ratio of hyperlinks to normal text within a web page would be more suitable to detect "Index" pages than searching for words that introduce a list of hyperlinks. We set an arbitrary threshold between the two groups based on the interestingness of discovered relations and a minimal achieved F-Measure of 0.25. In respect to to the threshold, the group of word-describable genres contains 15 genres, the analysis of which (except "Blog") is presented in the following subsection. The second group contains five genres ("Commercial/promotional", "Index", "Gateway", "Content delivery", "Informative"), which are not discussed any further.

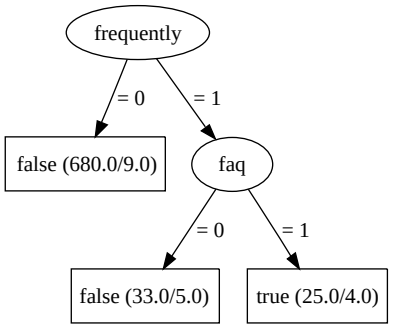
6.4.2 Word-Describable Genres

This subsection presents the credible models and relations that describe 14 word-describable genres.

FAQ. A credible tree for genre "FAQ" (frequently asked questions) is presented in Figure 6.6. The tree shows that this genre is best described with the combination of two words: "frequently" and "faq".

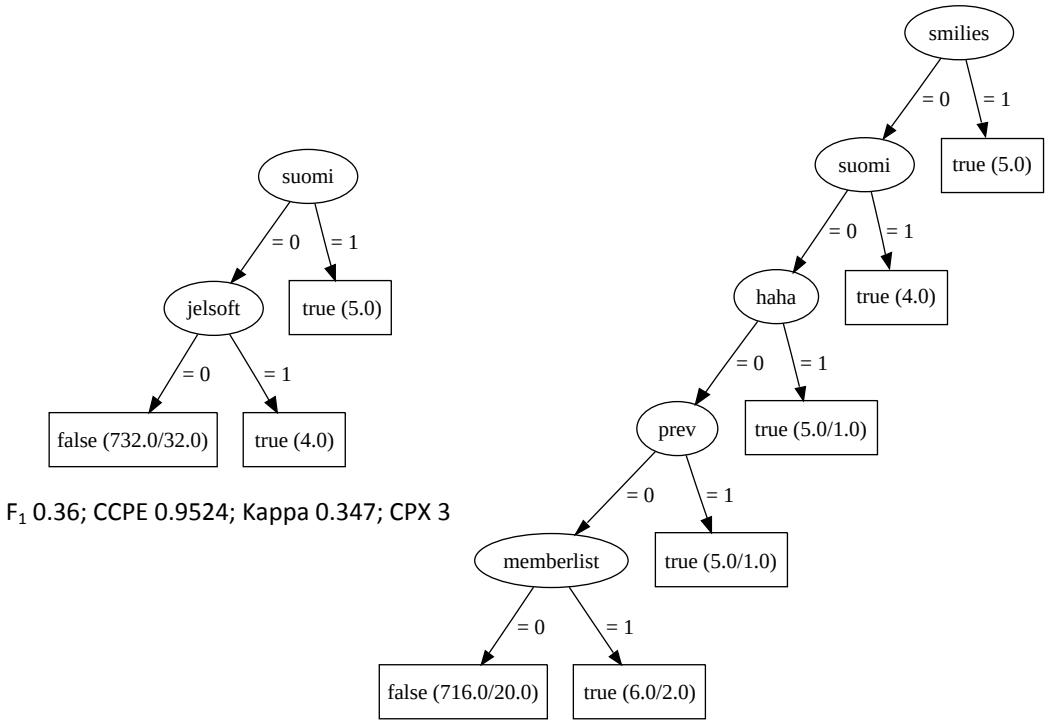
Community. The genre "Community" assumes such pages as forums, mailing list archives and portals with user-generated content (e.g., Wikipedia). It is best described with two credible trees presented in Figure 6.7. The relations within the trees state that when one of the six words (suomi, jelsoft, smilies, haha, prev, memberlist) appears within a page, the page belongs to the genre "Community". The explanation of the words is the following. "Suomi" is the official language of Finland. It appears within the Wikipedia pages, as one of the languages in which the article is written. One could expect that the word "languages" would appear instead. However, within the analysis, this specific

language appeared as a stronger representative of the list of languages than the word "languages". Then, there is a subset of three words, typical for forums: jelsoft, smilies and memberlist. "Jelsoft" represents a company that produced vBulletin – a commercial internet forum software. The word appears within the copyright at the bottom of the forum page. The words "Smilies" appears in the contexts such as "Smilies are on", meaning that the graphical representation of smilies will be used within the forum posts. "Memberlist" denotes a list of the forum's members. The word "prev" appears within the mailing list archives. It is an abbreviation of the word "previous", which is used for the purpose of navigation between the messages. The word "haha" is a part of informal communication characteristic for "Community" pages (e.g., "haha let's try that again").



F_1 0.7; CCPE 0.9377; Kappa 0.6877; CPX 3

Figure 6.6: The credible tree constructed for the genre "FAQ".

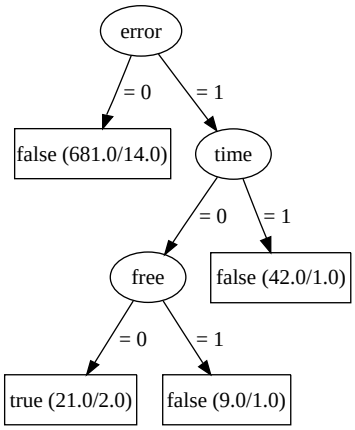


F_1 0.36; CCPE 0.9524; Kappa 0.347; CPX 3

F_1 0.5; CCPE 0.9498; Kappa 0.4818; CPX 6

Figure 6.7: The credible tree constructed for the genre "Community".

Error message. This genre represents the pages containing custom HTTP errors and non-HTTP errors, as well as those pages stating that the desired content is not available any more. The pages are best described with the word "error", while at the same time they should not contain the words "time" and "free" (see the credible tree in Figure 6.8).

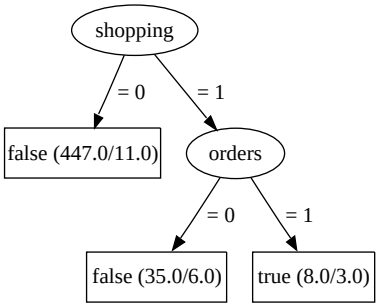


F_1 0.6786; CCPE 0.911; Kappa 0.667; CPX 4

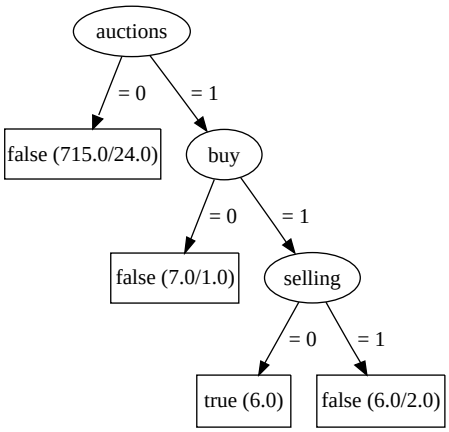
Figure 6.8: The credible tree constructed for the genre "Error message".

The word "free" frequently appears within the offers to claim a free service or product, while the word "time" often denotes time zones. "Error Message" pages mostly contain short messages without any additional content such as advertisements, which can explain the exclusion of the two words.

Shopping. This genre is composed of pages with online stores, classified advertisements, price comparators and price-lists. The pages are best represented by the two credible trees presented in Figure 6.9, which contain words that are straightforward representatives of a "Shopping" page: shopping, orders, auctions, buy and selling. The second tree is a good representative of such pages as ebay, where one of the ways of buying products is through auctions.



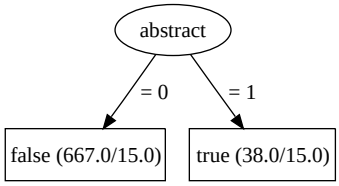
F_1 0.3556; CCPE 0.6306; Kappa 0.3397; CPX 3



F_1 0.2927; CCPE 0.9449; Kappa 0.2801; CPX 4

Figure 6.9: The credible tree constructed for the genre "Shopping".

While one can assume that words such as price, bid, shipping or chart may be good representatives of "Shopping" pages, the analysis showed that these words are of less importance. Within the "Shopping" pages these words usually appear as part of phrases: best price, add to chart, compare prices, return policy, shipping information, order status, etc. Considering that the attribute selection method used within the data preprocessing stage (see Section 6.2) did not account for interactions between the attributes, some words from the phrases were eliminated and could not be tested with the help of the HMDM. We believe that this is the reason that the credible trees achieved an F-Measure of only 0.318.

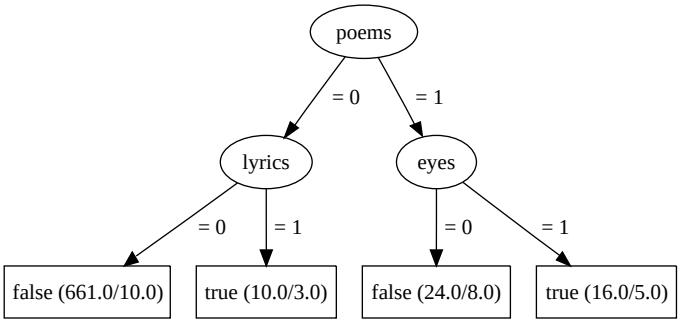


F_1 0.6053; CCPE 0.9574; Kappa 0.5828; CPX 2

Figure 6.10: The credible tree constructed for the genre "Scientific".

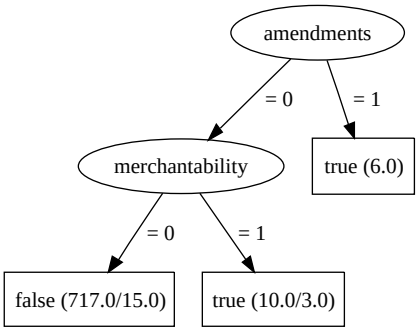
Scientific. "Scientific" pages contain scientific papers, theses, lecture notes for a specialized audience and scientific books. This category is easy explainable (see the credible tree in Figure 6.10): if the page contains the word "abstract" than it is a "Scientific" page. We experimented with different combinations of words, which represent parts of scientific documents (e.g., introduction, background, references, conclusions). However, it seems that the most common to all "Scientific" pages is that they contain an abstract.

Poetry. This category contains two types of pages, one with poems and another with lyrics. The credible tree presented in Figure 6.11 reflects these two subcategories. In other words, the page belongs to the genre "Poetry" if it does not contain the word "poems", but contains the word "lyrics" (left subtree), or if it contains a combination of the words "poems" and "eyes" (right subtree). The second relation is interesting, since both words on their own form rather weak relations, while the combination brings considerable improvement in quality. Further analysis justified the second relation by showing that many poems indeed mentioned eyes in one context or another.



F_1 0.5806; CCPE 0.9378; Kappa 0.562; CPX 4

Figure 6.11: The credible tree constructed for the genre "Poetry".

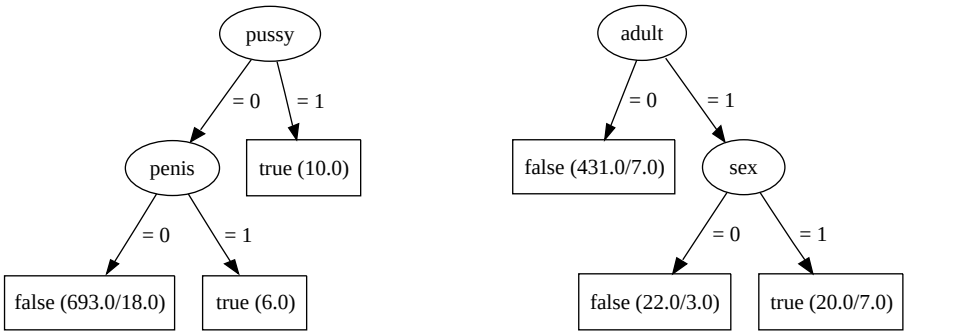


F_1 0.5909; CCPE 0.9724; Kappa 0.5792; CPX 3

Figure 6.12: The credible tree constructed for the genre "Official".

Official. "Official" pages contain legal materials, official reports and rules. The credible tree in Figure 6.12 differentiates between two types of "Official" pages. The first

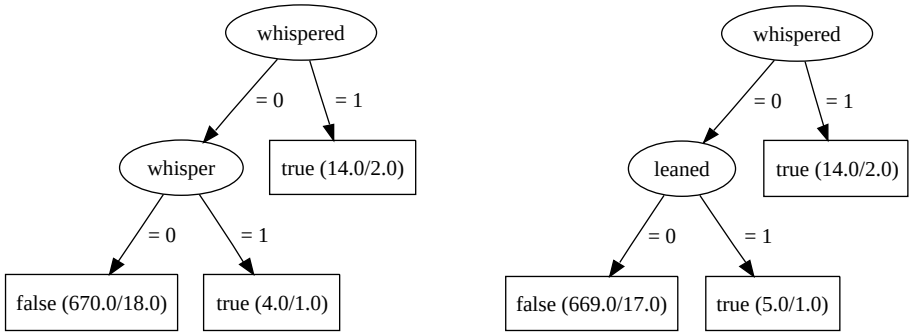
contains pages that report about the terms of use for a product or a service, where the word "merchantability" denotes the product or service ready to be sold. In contrast, the second type of legal documents is characterized as those containing amendments, i.e., minor changes (e.g., legal acts).



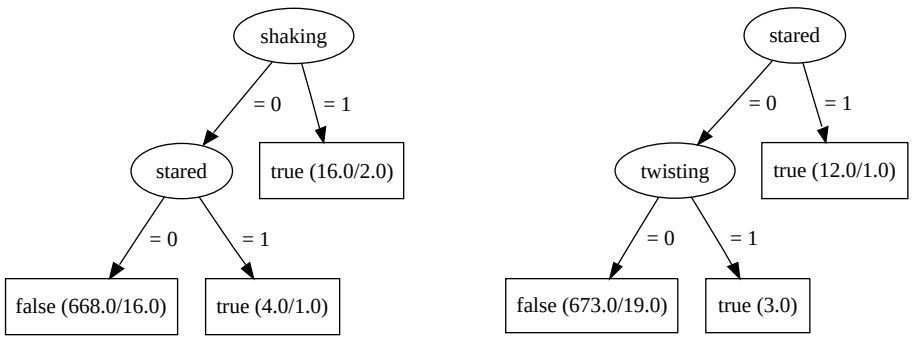
F_1 0.64; CCPE 0.9725; Kappa 0.6286; CPX 3 F_1 0.6364; CCPE 0.9237; Kappa 0.6186; CPX 3

Figure 6.13: The credible tree constructed for the genre "Pornographic".

Pornographic. This category contains pages with pornographic stories, pictures and videos, as well as stores with sex toys and pornographic magazines. The genre is described with two credible trees stating that a "Pornographic" page is characterized either by the appearance of one of the two words "pussy" or "penis", or by the appearance of a combination of the words "adult" and "sex".



F_1 0.5882; CCPE 0.966; Kappa 0.5738; CPX 3 F_1 0.56; CCPE 0.9675; Kappa 0.5452; CPX 3



F_1 0.6415; CCPE 0.9689; Kappa 0.628; CPX 3 F_1 0.5833; CCPE 0.968; Kappa 0.5705; CPX 3

Figure 6.14: The credible tree constructed for the genre "Prose fiction".

Prose fiction. "Prose fiction" pages contain stories, which are best described with one or several of the following words: whispered, whisper, leaned, shaking, stared, twisting (see the credible trees in Figure 6.14). "Whispered" and "whisper" denote a manner of communication used by a character in a story (e.g., "he whispered to her" or "she heard her

mother's voice whisper"). "Leaned" and "stared" describe a non-verbal communication of characters (e.g., "he leaned toward her", "stood up and leaned over his table", "she stared out over the town", "she stared at him"). Similarly, "shaking" describes a non-verbal communication (e.g., "anxiously shaking Steve's hand") or a character's property (e.g., "her shaking hands"). "Twisting" appears in a combination with different objects (e.g., twisting of a spaceship after the crash in science-fiction stories).

Children's. This category is best described with the two trees presented in Figure 6.15. Although the first tree is of a quality lower than the second (measured by cross-validation within the HMDM), we chose that tree, since it was the best observed tree composed of words that actually describe the category. In contrast, the second tree mostly contains words that do not appear within the "Children's" pages.

The "Children's" page contains one of the following words: coloring, printouts or kids. "Coloring" appears in the context of coloring books or coloring pages, representing colorless drawings ready to be printed. Similarly, "printouts" also refer to drawings ready to be printed, although they are not always intended for coloring. These drawings are frequently created with an intention to educate, like, e.g., a drawing of a human with marked body parts. "Children's" pages often directly refer to kids, which explains an appearance of the words "kids" (e.g., "kids in action", "activities for kids", "Hey Kids what do you think about ..." or "Fun Stuff for Kids").

The difficulties in learning this category lie in a considerable amount of graphical elements within the "Children's" pages. A lot of words is represented as graphical elements and are not computer-readable.

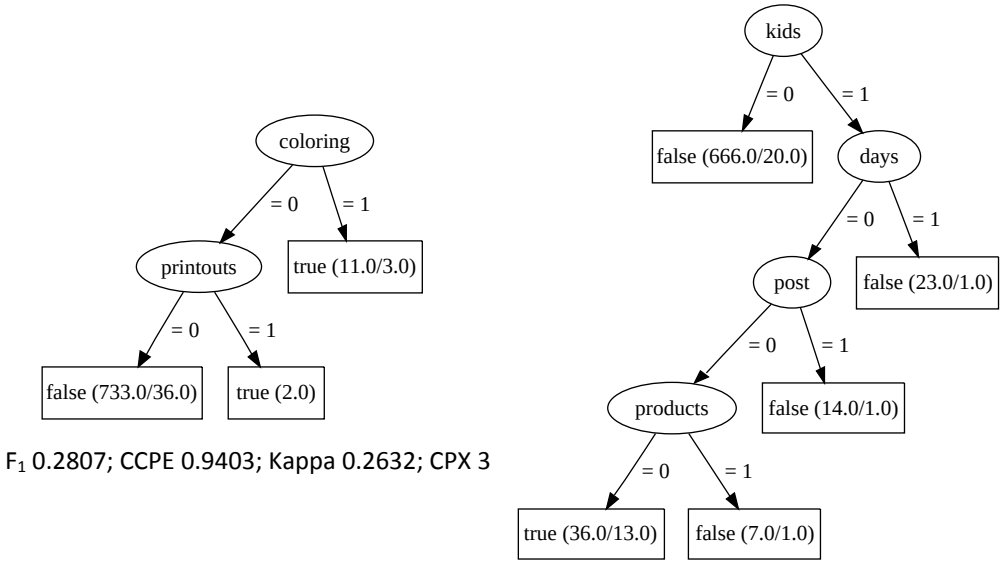


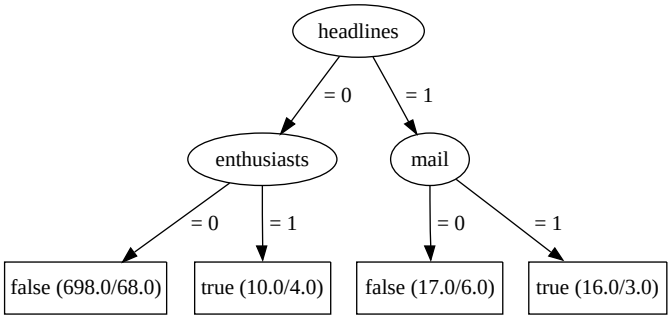
Figure 6.15: The credible tree constructed for the genre "Childrens'".

Journalistic. This category contains pages with news, reportages, editorials, interviews and reviews. For this category it was difficult to find high-quality trees that did not contain the names of journalists or persons addressed in articles (e.g., ridley, jenni, patterson, peggy), or names of places (e.g., Guantanamo). Therefore, between the meaningful trees we selected those that had the highest quality measures. The best trees are presented in Figure 6.16.

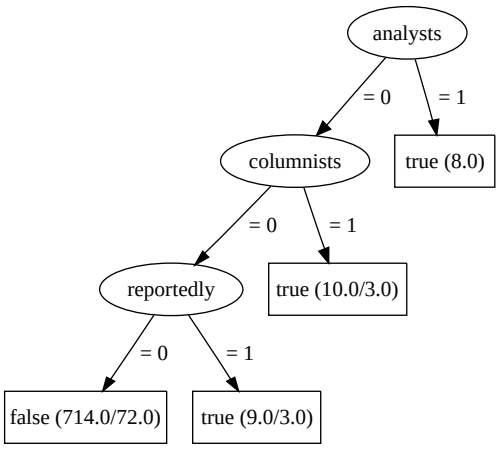
The first tree describes two types of "Journalistic" pages. The first type, represented with the left subtree, assumes those pages that mostly report about leisure activities. The pages contain the word "enthusiasts", which denotes persons enthusiastic about a

topic of article (e.g., "Craft enthusiast began arriving at 10am.", "The club was founded by a group of enthusiasts."). In contrast, the second type, represented with the right subtree, assumes such pages that report about recent activities (e.g., news, sport, weather, entertainment). Those pages have a characteristic structure, composed of an article and supporting menus, which contain links to other "headlines" and different means for the user to receive the latest news over an e-"mail" (e.g., "subscribe to e-mail newsletters").

The second tree differentiates between three types of "Journalistic" pages. The pages that discuss economic issues are characterized with the word "analysts", which mostly refers to the statements made by expert analysts. Second, the pages that report about recent activities often contain a link to other articles written by "columnists" (e.g., the link "Our columnists"). Finally, the bottommost subtree describes the yellow-press type of pages. The word "reportedly" appears in the context where a journalist retells what he/she heard that other person said or did (e.g., "The blonde singer was reportedly smitten with the shop's jewellery.").



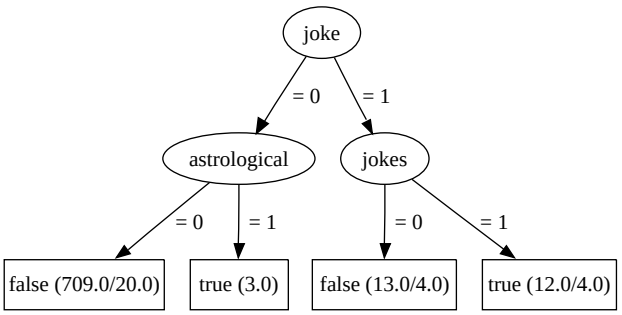
F_1 0.2906; CCPE 0.8625; Kappa 0.2521; CPX 4



F_1 0.322; CCPE 0.8782; Kappa 0.284; CPX 4

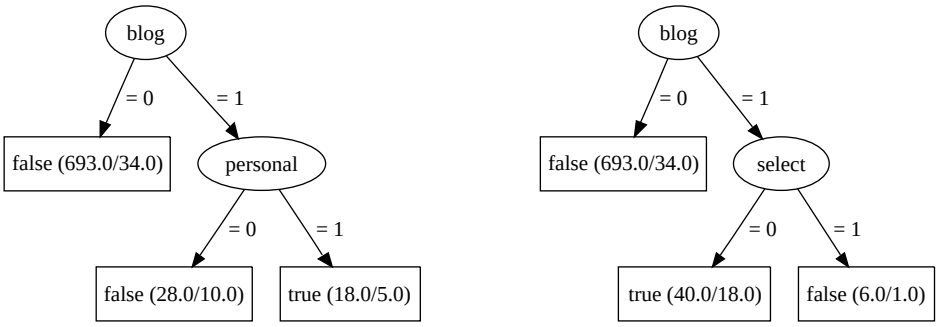
Figure 6.16: The credible tree constructed for the genre "Journalistic".

Entertainment. This category includes pages containing jokes, puzzles, horoscopes and games. Puzzles and games are mostly represented in graphical form or as animations. Since text within the graphics and animations is not computer-readable, the two subcategories cannot be described with words, which explains the F-Measure of 0.36. The other two categories – jokes and horoscopes – are described with two relations within the credible tree in Figure 6.17. The left subtree describes pages with horoscopes, which are characterized by the appearance of the word "astrological". The right subtree describes pages with jokes, which contain both words "joke" and "jokes".



F_1 0.375; CCPE 0.9435; Kappa 0.3585; CPX 4

Figure 6.17: The credible tree constructed for the genre "Entertainment".

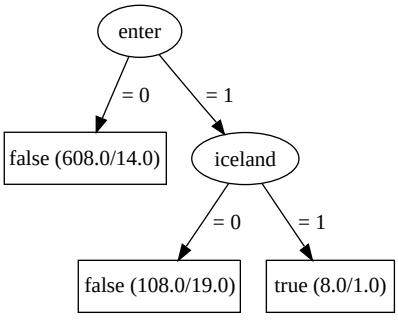


F_1 0.3467; CCPE 0.9081; Kappa 0.3215; CPX 3 F_1 0.4211; CCPE 0.8295; Kappa 0.383; CPX 3

Figure 6.18: The credible tree constructed for the genre "Personal".

Personal. This category contains pages made by individuals with the purpose to present their work or express their interests and opinions. One can expect that words which indicate a statement of a personal opinion (e.g., think, believe, feel, opinion – "in my opinion, ..."), would emerge as credible. However, only those words that describe a subgroup of "Blog" pages with a personal content emerged as credible.

The two credible trees are presented in Figure 6.18. They describe a "Personal" page as those containing words "blog" and "personal", while at the same time the page does not contain the word "select". Both genre-specific words describe blogs, since the word "personal" represents one of the blog categories. Since the two trees do not describe other subgroups, such as personal homepages, the F-Measure is low (0.255).



F_1 0.2917; CCPE 0.8403; Kappa 0.2784; CPX 3

Figure 6.19: The credible tree constructed for the genre "User input".

User input. This genre contains forms and surveys. The most dominant words that appeared within the analysis were the names of countries extracted from the field requiring from the user to select one of the countries. Considering that the same set of countries

appeared throughout the pages containing the field “country”, a single tree composed of an arbitrary country is enough to represent the type of pages containing this field. We also examined the role of the word “country” as a meaningful substitute for the names of countries, however, this word was not even selected between the top 500 words during the automatic attribute selection procedure. The second word “enter” appeared in the contexts as “enter keyword and item number” or “enter information”, which represent as straightforward representative of the “User input” pages.

In summary, in this chapter we presented an application of the HMDM method for the construction of credible predictive models. On the task of AWGI, we showed that the credible models outperform the models constructed by automatic ML method in both the meaning and quality. Furthermore, we showed that the majority of genres (15 of 20) from the 20-Genre corpus are describable with meaningful genre-specific content words, which are presented for each of the 15 genres.

7 Evaluation

With the help of a user study we aimed to determine: Q1) whether users recognize the less-credible relations based on a single tree and its quality? and Q2) do they agree with the improvements proposed by the HMDM method?

We chose a set of trees from the experiment using the HMDM and the R&D data (Section 5.2) and organized them in the form of a paper-based questionnaire. In total, 22 users participated in the study, all of which had prior knowledge of the decision trees.

The experiment was conducted one participant at a time with a facilitator interacting with the participant. Since the participants were not familiar with all the attributes, the facilitator helped by answering technical questions. The facilitator, however, did not influence in any way the participant's choices and did not indicate which algorithm was used to construct the presented decision trees. The participants were encouraged to express any comment considering the task. The experiment was performed page by page in one pass with no returning back or looking in advance. The data set, the questionnaire and the table with all the answers are available at http://dis.ijs.si/Vedrana/user-study.htm. In addition, the questionnaire is presented in Appendix C. It contains additional questions that are not discussed in this chapter.

The questionnaire is composed of two parts. Altogether, there are four trees in it, one presented in Figure 1.1 and three in Figure 7.1a-c. The first part of the questionnaire corresponds to Q1 and Figure 1.1, followed by questions intended to determine whether the participants find the presented tree reasonable. We use the term "reasonable" without an additional explanation to observe the criteria on which the participants base their decisions.

The first part of the questionnaire contains four questions with yes/no answers:

1. Does the tree sound reasonable or not?

2. Is the attribute in the root node (that the most important factor for the welfare of a country is the level of investment in R&D) reasonable or not?

3. Does the right subtree (starting with "Sector investing the most in R&D") present reasonable relations or not?

4. Does the left subtree (starting with "Sector employing the most researchers") present reasonable relations or not?

The answers are presented in Table 7.1. We divided the participants into two groups: first, those with the initially majority positive answers (64%); and second, those with the majority negative answers or an equal number of positive and negative answers (36%). The positive group typically stated that the initial tree and all of its parts are reasonable (86% of the positive group); while the negative group typically stated that the choice of a root node sounds reasonable, while the tree in general and one or both of its subtrees sound unreasonable (75% of the negative group). Some participants from the positive group commented that the structure of the initial tree was strange, e.g., branches in the

Table 7.1: The results of the user study.

	QUESTIONS						
PARTICIPANT	1	2	3	4	5	6	7
1	yes	yes	yes	yes	2,4	cont.	2,4,1,3
2	no	no	no	no	2,4	cont., CPX	2,3,4,1
3	no	yes	no	no	2,3,4	cont.	–
4	no	yes	no	no	4	cont.	4,3,2,1
5	yes	yes	yes	yes	4	C/K	2,3,4,1
6	no	yes	no	no	2	cont.	2,3,4,1
7	yes	yes	yes	yes	2,3,4	CPX	3,2,4,1
8	yes	yes	yes	no	2,4	ACC, cont.	2,4,3,1
9	yes	yes	yes	yes	3,4	cont.	3,2,4,1
10	yes	yes	yes	yes	4	cont.	2,3,4,1
11	yes	yes	no	yes	2,4	cont.	2,3,4,1
12	yes	yes	yes	yes	2,4	cont.	2,3,4,1
13	yes	yes	yes	yes	2	CPX, cont.	2,4,1,3
14	yes	yes	yes	yes	2,4	cont.	2,1,3,4
15	yes	yes	yes	yes	4	cont.	4,2,3,1
16	yes	yes	yes	yes	2	CPX, cont.	3,2,1,4
17	yes	yes	yes	yes	2,4	CPX	2,4,3,1
18	yes	yes	yes	yes	2,4	cont., CPX	4,2,3,1
19	no	yes	yes	no	3	CPX, C/K, cont.	3,4,1,2
20	no	yes	no	yes	2,3	CPX	3,2,1,4
21	yes	yes	no	no	2,4	cont., CPX	3,2,1,4
22	no	yes	no	no	2,3,4	cont., CPX	2,3,4,1
SUM	y:15	y:21	y:14	y:14	4:17/22	cont.:18/22	2,3,4,1:7
	n:7	n:1	n:8	n:8	2:16/22	CPX:10/22	3,2,1,4:3
					3:6/22	C/K:2/22	2,4,3,1:2
					∅:0/22	ACC:1/22	2,4,1,3:2
				3,2,4,1:2;	4,2,3,1:2;	2,1,3,4:1;	3,4,1,2:1; 4,3,2,1:1

right subtree with none of the examples included, the presence of a N/A branch that does not bear important information; however, they considered the tree as semantically meaningful and this prevailed in the positive decision. In contrast, the participants from the negative group based their decisions primarily on the structure.

In conclusion, the participants mainly perceived at least some problems within the tree. However, overall, 64% of them accepted the initial tree.

The second part corresponds to Q2 and reveals the improvements, followed by questions aimed to determine whether the participants accept those improvements. In the second part, three improvements to the initial tree ranked highly according to the quality criterion were presented to the participant: the tree in Figure 7.1a was denoted as the second tree (the first is the initial tree), the tree in Figure 7.1b as the third, and the tree in Figure 7.1c as the fourth. Each tree was supported by the quality measures and the differences in quality from the initial tree. The participant would obtain the same trees by clicking on credibility indicators of an added attributes graph from which the trees were extracted; however, the participants did not have access to the program, just to one sheet of paper at a time. Our aims were to understand whether the participants would accept the modified trees as more credible than the initial tree and based on which criteria.

Questions 5 and 6 are connected to question 1, and question 7 is connected to question 2:

5. Which of the additional three trees sounds more reasonable than the first one? (a multiple choice question – answers: none (∅), second, third, fourth)

6. What is your decision based upon? (a multiple choice question – answers: ACC, CPX, other measures: CCPE and/or Kappa (C/K), content of the tree (cont.)).

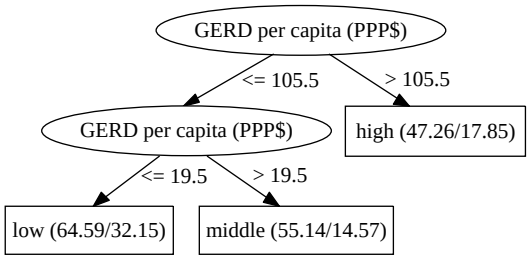
7. Write down the sequence of trees that persuaded you the most that the GERD attributes are the most important for the welfare of a country? E.g., 2, 3, 1, 4 means that 2 was the most persuasive and 4 the least.

All the participants stated that at least one of the three additional trees sounds more reasonable than the first one (question 5), indicating that the participants generally accept the improvements suggested by the HMDM method. Some 77% of the participants selected the tree in Figure 7.1c, 73% selected the tree in Figure 7.1a, and 27% selected the tree in Figure 7.1b. The answers to question 5 are also supported by the answers to question 7, where 64% of the participants ranked all the modified trees higher than the initial tree.

Figure 7.1a splits the "GERD per capita (PPP$)" into three intervals with higher values denoting better welfare. The tree in Figure 7.1b uses the "GERD per capita (PPP$)" and "GERD as % of GNI" to show similar relations as in Figure 7.1a with an exception for countries rich in natural resources. Figure 7.1c shows similar relations using the "GERD per capita (PPP$)" and "Applications for patents per researcher (HC)".

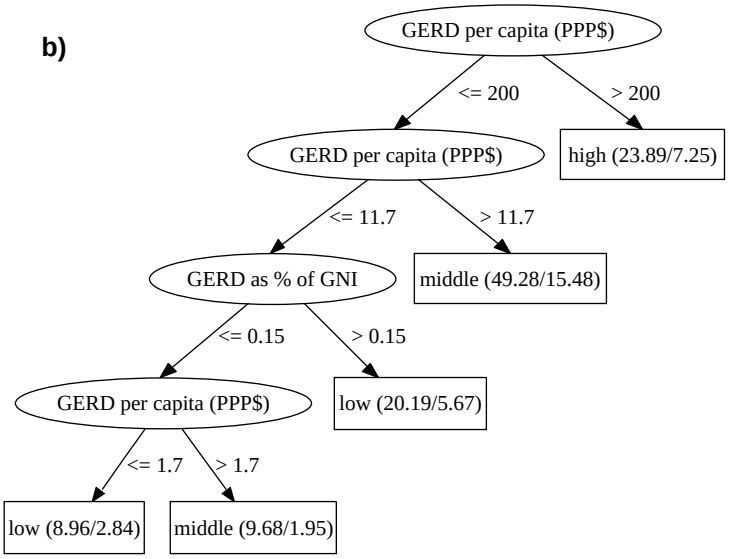
The most-selected criterion for choosing the modified trees as more reasonable than the initial tree is the improvement in the content of a tree (question 6). In total, 82% of the participants stated that their decision was based on the content, of which 61% based their decisions solely on the content. Some 45% stated that they preferred less complex trees, while only 14% stated that any of the quality estimates (ACC, CCPE/Kappa) played a part in their decision. The participants frequently stated that the differences in the qualities are small, which is the most probable reason for not selecting the quality as the relevant criterion. However, such behavior is desirable in the case of the HMDM method, since the system pre-selects the trees based on the quality and it is up to the user to

a)



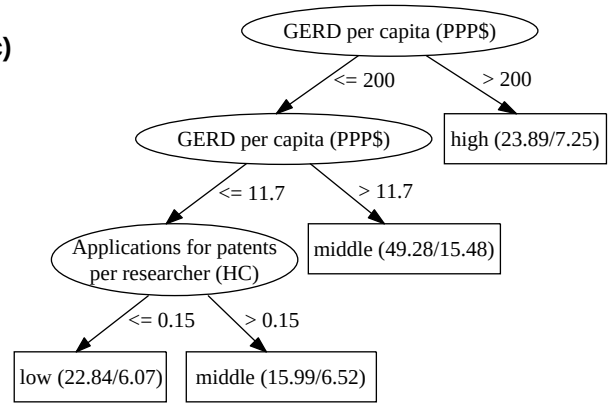
Accuracy : 63.47% (-1.2)
Kappa: 0.4023 (-0.0024)
Complexity : 3 relations (-8)
CCPE: 0.5685 (+0.1572)

b)



Accuracy : 66.47% (+1.8)
Kappa: 0.4314 (+0.0267)
Complexity : 5 relations (-6)
CCPE: 0.5287 (+0.1174)

c)



Accuracy : 66.47% (+1.8)
Kappa: 0.432 (+0.0273)
Complexity : 4 relations (-7)
CCPE: 0.5175 (+0.1062)

Figure 7.1: The improved trees presented within the questionnaire.

make the final decisions based primarily on the content. Additionally, the users saw only a couple of best trees with similar estimates and not the huge number of worse and much worse trees. In our experiments, the measures are welcome as a fast-elimination criterion for inferior candidates and later for providing additional information when comparing similar models.

In conclusion, the participants changed their opinion in favor of the modified trees, mostly based on the better estimated and more meaningful content of a tree, even though they had a chance to see only three best trees. It is assumed that the participants would be even more convinced had they had additional access to the program online.

8 Discussion and Conclusions

In this thesis we present a new method – Human-Machine Data Mining (HMDM). Its primary advantage is based on the interaction between the two most advanced information mechanisms: the brute force of computers enriched with DM and human insight and comprehension. The implemented interactive system constructs a set of models, in our experiments decision and regression trees, and explains the results to a human. The human leads the DM with the goal to find meaningful and high-quality models. In this way, meaningful models can be found of better quality than with "classical" DM approaches.

It is worth noting that the HMDM method is designed as an interactive rather than an automatic method for two reasons. First, the HMDM is designed to support the user in the learning process, by enabling him/her to interactively explore and learn about the domain of interest. Second, it is hard to formalize what is meaningful for the user. Therefore, the HMDM method is designed to combine the best of both worlds: formal measures of quality and informal knowledge and common sense provided by the user.

The method was demonstrated in two domains selected to answer the question as to how higher education and R&D influence the economic welfare of a country. To enable verification that the method is stable and that the results should not vary much depending on a particular human performing the HMDM, the data and results are accessible via the internet (see http://dis.ijs.si/Vedrana/economic-analysis.htm).

Furthermore, the HMDM was demonstrated on a complex task of learning predictive models for automatic web genre identification. Due to the elimination of less-credible relations, the credible models exhibited higher predictive performance than the models constructed with "classical" DM approaches. In addition, when the credible models could not be obtained, the HMDM provided an explanation of reasons.

However, the debate is open regarding which type of relation was indeed observed – X implies Y or Y implies X? Does more investment in education actually cause countries to progress faster or is it just a side-effect of the developed countries spending more on education? Although the trees and analyses in this paper do not indicate the type of relation, in our opinion it is highly unlikely that such a strong relation would not be mutual, acting in both directions. But to evaluate these relations quantitatively, other methodologies are more appropriate than HMDM.

In general, it depends on human ingenuity to accept or reject any conclusion, however, supported by statistics or another formal method. By observing not only one model in one DM setup, but thousands of them and giving an interactive tool to verify the hypotheses enabling the human mind to integrate conclusions from thousands of constructed transparent models, the summarized relations emerge as indeed credible.

The HMDM method was justified through the user study, showing that the users are often not able to detect weak subtrees in the automatically generated trees. When faced with better solutions provided by HMDM, all of the 22 participants realized the weaknesses in the default tree and accepted the better ones.

As part of the future work, five improvements of the HMDM seem interesting. First, in this thesis we used the HMDM method in combination with decision and regression

trees. However, we assume that the method is applicable to any other supervised DM method that produces a model in a human-understandable form. As part of the future work, it would be interesting to test the HMDM in combination with other DM methods. Second, the flexible quality criterion for ranking the models within the removed and added attributes graphs could be implemented. Flexible means that the user can attribute different weights to the selected quality measures. In this manner, the user will tune the algorithm to give higher ranks to those models he/she considers more credible. Third, the weights within quality-based ranking criterion may be learned from models marked as credible by the user. In contrast, learning of the meaning-based criterion from the credible models is problematic, since it is hard to formalize common sense. Fourth, we intend to improve the visualization of the removed and added attributes graphs. Finally, it seems to be interesting to implement and explore the interactive explanations of the model's subparts (e.g., a subtree within the decision tree) in a form similar to ensemble trees.

9 Acknowledgments

This thesis would not have been possible without the generous help and support of my colleagues and family. First of all, I would like to thank my supervisor prof. dr. Matjaž Gams, who has provided guidance, support, understanding and professional and personal assistance of the most valuable kind. I am also very thankful to prof. dr. Bogdan Filipič and prof. dr. Marko Bohanec, who monitored my progress and helped me to focus my research.

I have to thank to my colleagues from Department of Intelligent Systems, several of whom deserve special merits. Boštjan Kaluža shared with me many new valuable ideas and helped me to solve problems with software, especially problems with LaTeX, which emerged in the last stages of writing this thesis. Damjan Kužnar and Jana Krivec helped me in the early stages of the development of the HMDM method. The research on web genres presented in this thesis was built on the work done in collaboration with dr. Mitja Luštrek and Aleš Tavčar. I am very thankful to Sanja Kovač who helped me to conduct experiments with genres.

I am very thankful to my family, Dino, Antica and Marčelo, who believed in me and supported me all the way to the end. They helped me to regain the positive attitude in the toughest times of my study.

I also need to thank to my classmates and good friends Panče Panov, Ivica Slavkov and Dragi Kocev, who shared the same burden of PhD study and helped me make my everyday commitments much easier.

Last, but not the least, I am grateful to the Jožef Stefan Institute and the Slovene Human Resources Development and Scholarship Fund for providing me a scholarship, which made this thesis possible.

10 References

Becker, B.; Kohavi, R.; Sommerfield, D. Visualizing the Simple Bayesian Classifier. In: Fayyad, U.; Grinstein, G.; Wierse, A. (eds.) *Information Visualization in Data Mining and Knowledge Discovery*, 237–249 (Morgan Kaufman, San Francisco, CA, 2001).

Black, J.; Hashimzade, N.; Myles, G. *A Dictionary of Economics* (Oxford University Press, New York, 2009).

Bohanec, M.; Bratko, I. Trading Accuracy for Simplicity in Decision Trees. *Machine Learning* **15**(3):223–250 (1994).

Breiman, L. Random Forests. *Machine Learning* **45**(1):5–32 (2001).

Burrows, E.; Wallace, M. *Gotham: A History of New York City to 1898* (Oxford University Press, New York, 1999).

Cohn, D.; Ghahramani, Z.; Jordan, M. Active Learning with Statistical Models. *Journal of Artificial Intelligence Research* **4**:129–145 (1996).

Craven, M. *Extracting Comprehensible Models from Trained Neural Networks.* Ph.D. thesis, School of Computer Science, University of Wisconsin, Madison, WI (1996).

Culotta, A.; Kristjansson, T.; McCallum, A.; Viola, P. Corrective Feedback and Persistent Learning for Information Extraction. *Artificial Intelligence* **170**:1101–1122 (2006).

Davies, S. The Great Horse-Manure Crisis of 1894. *The Freeman. Ideas On Liberty* **54**(7) (2004).

Dekking, F.; Kraaikamp, C.; Lopuhaä, H.; Meester, L. *A Modern Introduction to Probability and Statistics: Understanding Why and How* (Springer Verlag, 2005).

Demšar, J.; Zupan, B.; Leban, G.; Curk, T. Orange: From Experimental Machine Learning to Interactive Data Mining. In: *Proceedings of 8th European Conference on Principles and Practice of Knowledge Discovery in Databases*, 537–539 (Pisa, Italy, 2004).

Dietterich, T. Approximate Statistical Tests for Comparing Supervised Classification Learning Algorithms. *Neural Computation* **10**(7):1895–1923 (1998).

Fails, J.; Olsen, D. Interactive Machine Learning. In: *Proceedings of the 8th International Conference on Intelligent User Interfaces – IUI 2003*, 39–45 (Miami, FL, 2003).

Filipič, B.; Urbančič, T.; Križman, V. A Combined Machine Learning and Genetic Algorithm Approach to Controller Design. *Engineering Applications of Artificial Intelligence* **12**(4):401–409 (1999).

Fisher, R.; Tippett, L. Limiting Forms of the Frequency Distribution of the Largest and Smallest Member of a Sample. *Proceedings of the Cambridge Philosophical Society* **24**:180–190 (1928).

Furman, J.; Porter, M.; Stern, S. The Determinants of National Innovative Capacity. *Research Policy* **31**:899–933 (2002).

Gams, M.; Krivec, J. Demographic Analysis of Fertility Using Data Mining Tools. *Informatica* **32**:147–156 (2008).

Gylfason, T. Natural Resources, Education, and Economic Development. *European Economic Review* **45**:847–859 (2001).

Huang, Y.; Mitchell, T. Text Clustering with Extended User Feedback. In: *Proceedings of the 29th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, 413–420 (Seattle, WA, 2006).

Jakulin, A. *Machine Learning Based on Attribute Interaction*. Ph.D. thesis, Faculty of Computer and Information Science, University of Ljubljana (2005).

Jensen, D.; Cohen, P. Multiple Comparisons in Induction Algorithms. *Machine Learning* **38**(3):309–338 (2000).

John, G. *Enhancements to the Data Mining Process*. Ph.D. thesis, Computer Science Department, Stanford University, Stanford, CA (1997).

Karlgren, J. *Stylistic Experiments for Information Retrieval*. Ph.D. thesis, Stockholm University (2000).

Kearns, M.; Vazirani, U. *An Introduction to Computational Learning Theory* (The MIT Press, Cambridge, MA, 1994).

Keller, K. Investment in Primary, Secondary, and Higher Education and the Effects on Economic Growth. *Contemporary Economic Policy* **24**(1):18–34 (2006).

Kohavi, R. A Study of Cross-Validation and Bootstrap for Accuracy Estimation and Model Selection. In: *Proceedings of International Joint Conference on Artificial Intelligence – IJCAI 1995*, 1137–1145 (Montréal, Québec, Canada, 1995).

Kohavi, R.; John, G. Wrappers for Feature Subset Selection. *Artificial Intelligence* **97**(1–2):273–324 (1997).

Kononenko, I. Inductive and Bayesian Learning in Medical Diagnosis. *Applied Artificial Intelligence* **7**(4):317–337 (1993).

Kulesza, T.; Wong, W.-K.; Stumpf, S.; Perona, S.; White, R.; Burnett, M.; Oberst, I.; Ko, A. Fixing the Program My Computer Learned: Barriers for End Users, Challenges for the Machine. In: *Proceedings of the 13th International Conference on Intelligent User Interfaces – IUI 2009*, 187–196 (Sanibel Island, FL, 2009).

Lacave, C.; Diez, F. A Review of Explanation Methods for Bayesian Networks. *Knowledge Engineering Review* **17**(2):107–127 (2002).

Langley, P. *Elements of Machine Learning* (Morgan Kaufmann, 1996).

Lewis, D.; Yang, Y.; Rose, T.; Li, F. Rcv1: A New Benchmark Collection for Text Categorization Research. *The Journal of Machine Learning Research* **5**:361–397 (2004).

Luštrek, M.; Vidulin, V.; Gams, M. Genres for Web Page Classification. In: *Proceedings of the International Conference on Advances in the Internet, Processing, Systems, and Interdisciplinary Research – VIPSI 2007* (2007).

MacKay, D. Information-Based Objective Functions for Active Data Selection. *Neural Computation* **4**(4):590–604 (1992).

Mardia, K.; Kent, J.; Bibby, J. *Multivariate Analysis* (Academic Press, London, 1979).

McShane, C.; Tarr, J. *The Horse in the City: Living Machines in the Nineteenth Century* (The Johns Hopkins University Press, 2007).

Melville, P.; Yang, S.; Saar-Tsechansky, M.; Mooney, R. Active Learning for Probability Estimation using Jensen-Shannon Divergence. In: *Proceedings of the 16th European Conference on Machine Learning – ECML 2005*, 268–279 (Porto, Portugal, 2005).

Mitchell, T. *Machine Learning* (McGraw-Hill, 1997).

Možina, M.; Demšar, J.; Kattan, M.; Zupan, B. Nomograms for Visualization of Naïve Bayesian Classifier. In: *Proceedings of 8th European Conference on Principles and Practice of Knowledge Discovery in Databases*, 337–348 (Pisa, IT, 2004).

Možina, M.; Demšar, J.; Žabkar, J.; Bratko, I. Why is Rule Learning Optimistic and How to Correct It. In: Fürnkranz, J.; Scheffer, T.; Spiliopoulou, M. (eds.) *Machine Learning: ECML 2006*, 330–340 (Springer, Berlin, 2006).

Nguyen, T.; Ho, T.; Shimodaira, H. A Visualization Tool for Interactive Learning of Large Decision Trees. In: *Proceedings of the 12th IEEE International Conference on Tools with Artificial Intelligence – ICTAI 2000*, 28–35 (Vancouver, BA, Canada, 2000).

OECD. Frascati Manual: Proposed Standard Practice for Surveys on Research and Experimental Development (2002).

Osei-Bryson, K. Evaluation of Decision Trees: A Multi-Criteria Approach. *Computers & Operations Research* **31**:1933–1945 (2004).

Poulin, B.; Eisner, R.; Szafron, D.; Lu, P.; Greiner, R.; Wishart, D.; Fyshe, A.; Pearcy, B.; MacDowell, C.; Anvik, J. Visual Explanation of Evidence with Additive Classifiers. In: *Proceedings of the Conference on Innovative Applications of Artificial Intelligence – IAAI 2006* (Boston, MA, 2006).

Quinlan, J. Learning with Continuous Classes. In: *Proceedings of 5th Australian Joint Conference on AI*, 343–348 (Singapore, 1992).

Quinlan, J. *C4.5: Programs for Machine Learning* (Morgan Kaufmann, San Mateo, CA, 1993).

Rehm, G.; Santini, M.; Mehler, A.; Braslavski, P.; Gleim, R.; Stubbe, A.; Symonenko, S.; Tavosanis, M.; Vidulin, V. Towards a Reference Corpus of Web Genres for the Evaluation of Genre Identification Systems. In: *Proceedings of the 6th International Language Resources and Evaluation – LREC 2008*, 351–358 (2008).

Salton, G.; Buckley, C. Term-Weighting Approaches in Automatic Text Retrieval. *Information Processing & Management* **24**(5):513–523 (1988).

Santini, M. *Automatic Identification of Genre in Web Pages*. Ph.D. thesis, University of Brighton (2007).

Sebastiani, F. Machine Learning in Automated Text Categorization. *ACM computing surveys (CSUR)* **34**(1):1–47 (2002).

Shepherd, M.; Watters, C.; Kennedy, A. Cybergenre: Automatic Identification of Home Pages on the Web. *Journal of Web Engineering* **3**(3–4):236–251 (2004).

Shilman, M.; Tan, D.; Simard, P. CueTIP: A Mixed-Initiative Interface for Correcting Handwriting Errors. In: *Proceedings of the 19th Annual ACM Symposium on User Interface Software and Technology – UIST 2006*, 323–332 (Montreux, Switzerland, 2006).

Stumpf, S.; Rajaram, V.; Li, L.; Wong, W.; Burnett, M.; Dietterich, T.; Sullivan, E.; Herlocker, J. Interacting Meaningfully with Machine Learning Systems: Three Experiments. *International Journal of Human Computer Studies* **67**(8):639–662 (2009).

Tong, S.; Koller, D. Support Vector Machine Active Learning with Applications to Text Classification. *Journal of Machine Learning Research* **2**:45–66 (2002).

Towell, G.; Shavlik, J. Extracting Refined Rules from Knowledge-Based Neural Networks. *Machine Learning* **13**(1):71–101 (1993).

Tsoumakas, G.; Katakis, I.; Vlahavas, I. Mining Multi-Label Data. In: Maimon, O.; Rokach, L. (eds.) *Data Mining and Knowledge Discovery Handbook*, 667–685 (Springer US, 2010).

UNESCO. Manual for Statistics on Scientific and Technological Activities (1984).

UNESCO. ISCED 1997 – International Standard Classification of Education (2006).

Varsakelis, N. Education, Political Institutions and Innovative Activity: A Cross-Country Empirical Investigation. *Research Policy* **35**(7):1083–1090 (2006).

Vidulin, V. Constructivist Learning Theory as a Link Between Artificial Neural Networks and Intelligent Tutoring Systems. *Organizacija* **39**(2):154–156 (2006).

Vidulin, V. Problem Transformation Methods for Multi-Genre Web Pages Classification. In: *Proceedings of the 12th International Multiconference Information Society – IS 2009*, 136–139 (2009).

Vidulin, V.; Filipič, B. Visualization of a Simple Genetic Algorithm for Pedagogical Purposes. In: *Proceedings of the 15th International Electrotechnical and Computer Science Conference – ERK 2006*, 99–102 (2006).

Vidulin, V.; Gams, M. Analyzing the Impact of Investment in Education and R&D on Economic Welfare with Data Mining. *Electrotechnical Review* **73**(5):285–290 (2006a).

Vidulin, V.; Gams, M. Impact of Investment in Education and R&D on Economic Growth. In: *Proceedings of the 15th International Electrotechnical and Computer Science Conference – ERK 2006*, 129–132 (2006b).

Vidulin, V.; Gams, M. The Impact of High Level Knowledge on Economic Welfare. In: *Proceedings of the 10th International Multiconference Information Society – IS 2007*, 107–110 (2007).

Vidulin, V.; Gams, M. Impacts of Education and R&D on Economy: Analysis by Data Mining Techniques. In: *Proceedings of the International Conference on Advances in the Internet, Processing, Systems, and Interdisciplinary Research – VIPSI 2008* (2008a).

Vidulin, V.; Gams, M. Is Science Important for Economic Welfare? In: *Proceedings of the 11th International Multiconference Information Society – IS 2008*, 41–44 (2008b).

Vidulin, V.; Gams, M. Multi-Label Classification of Web Genres. In: *Proceedings of the 18th International Electrotechnical and Computer Science Conference – ERK 2009*, 179–182 (2009).

Vidulin, V.; Gams, M. Searching for Meaningful Models in Macroeconomic Domain. In: *Proceedings of the 13th International Multiconference Information Society – IS 2010*, 94–97 (2010).

Vidulin, V.; Gams, M. Impact of High-Level Knowledge on Economic Welfare through Interactive Data Mining. *Applied Artificial Intelligence* **25**(4):267–291 (2011).

Vidulin, V.; Luštrek, M.; Gams, M. Comparison of the Performance of Genre Classifiers Trained by Different Machine Learning Algorithms. In: *Proceedings of the 9th International Multiconference Information Society – IS 2006*, 140–143 (2006).

Vidulin, V.; Luštrek, M.; Gams, M. Evaluation of Different Approaches to Training a Genre Classifier. In: *Proceedings of the 2007 International Conference on Artificial Intelligence and Pattern Recognition – AIPR 2007*, 515–520 (2007a).

Vidulin, V.; Luštrek, M.; Gams, M. Training a Genre Classifier for Automatic Classification of Web Pages. *Journal of Computing and Information Technology* **15**(4):305–311 (2007b).

Vidulin, V.; Luštrek, M.; Gams, M. Training the Genre Classifier for Automatic Classification of Web Pages. In: *Proceedings of the 29th International Conference on Information Technology Interfaces – ITI 2007*, 93–98 (2007c).

Vidulin, V.; Luštrek, M.; Gams, M. Using Genres to Improve Search Engines. In: *Towards Genre-Enable Search Engines: the Impact of Natural Language Processing - Proceedings of the International Workshop*, 45–51 (2007d).

Vidulin, V.; Luštrek, M.; Gams, M. Multi-Label Approaches to Web Genre Identification. *Journal for Language Technology and Computational Linguistics* **24**(1):93–110 (2009).

Štrumbelj, E.; Kononenko, I.; Robnik Šikonja, M. Explaining Instance Classifications with Interactions of Subsets of Feature Values. *Data & Knowledge Engineering* **68**(10):886–904 (2009).

Žnidaršič, M.; Bohanec, M. Automatic Revision of Qualitative Multi-Attribute Decision Models. *Foundations of Computing and Decision Sciences* **32**(4):315–326 (2007).

Ware, M.; Frank, E.; Holmes, G.; Hall, M.; Witten, I. Interactive Machine Learning: Letting Users Build Classifiers. *International Journal of Human-Computer Studies* **55**:281–292 (2001).

Witten, I.; Frank, E. *Data Mining: Practical Machine Learning Tools and Techniques* (Morgan Kaufmann, San Francisco, CA, 2005).

Xiao, B.; Wang, C.; Dai, R. Metasynthetic Approach for Handwritten Chinese Character Recognition. *International Journal of Information Technology and Decision Making* **1**(4):621–633 (2002).

Zhao, Y. *Interactive Data Mining* (VDM Verlag, 2008).

Zhao, Y.; Yao, Y. Interactive Classification Using a Granule Network. In: *Proceedings of the 4th IEEE Conference on Cognitive Informatics – ICCI 2005*, 250–259 (Irvine, CA, 2005).

Zhao, Y.; Yao, Y. On Interactive Data Mining. In: Wang, J. (ed.) *Encyclopedia of Data Warehousing and Mining*, 1085–1090 (Idea Group, London, 2008).

List of Figures

List of Tables

List of Algorithms

Appendix A: Higher Education Attributes

This appendix explains the higher education attributes used in the analyses presented in Section 5.1. The list of attributes is preceded by an explanation of general terms and categories. In the analysis we followed the definitions and classifications from the following sources: ISCED 97 (UNESCO, 2006) and UNESCO Institute for Statistics Online Education Glossary (http://www.uis.unesco.org/glossary).

A.1 International Standard Classification of Education

International Standard Classification of Education (ISCED) is a classification scheme used for the classification of educational programs into internationally comparable levels. The higher education programs are divided into two categories: first, ISCED 5 denotes the programs that belong to the first stage of higher education; second, ISCED 6 represents the programs that lead to the award of an advanced research qualification. The ISCED 5 programs are further divided into categories 5A and 5B. The 5A includes theoretically-based programs, preparing students for the ISCED 6 level programs or for high-skills professions. In contrast, the 5B includes practically-oriented programs that prepare students for employment in a particular occupation.

A.2 Expenditures on Education

Expenditures are coarsely divided into current and capital expenditures. Current expenditures include expenditures for goods and services consumed within the current year, for example, for staff salaries, pensions and benefits, contracted or purchased services, books and teaching materials. Capital expenditures include expenditures for assets that last longer than one year, for instance, construction, renovation and major repairs to buildings, and purchase of heavy equipment or vehicles.

A.3 Attributes

Attributes are divided into four groups that represent:

1. General characteristics of the higher education sector
2. Expenditures related to the higher education sector
3. Expenditures related to all levels of education
4. The previous three groups that are combined to form a set of constructed attributes

Attribute	Description
[**GER-Total**] Gross enrolment ratio. ISCED 5 and 6. Total	Gross enrolment ratio represents the number of students enrolled in higher education, regardless of their age, expressed as a percentage of the population in the five-year age group following on from leaving secondary school. It shows a general level of participation in higher education.
[**GER-Male**] Gross enrolment ratio. ISCED 5 and 6. Male	
[**GER-Female**] Gross enrolment ratio. ISCED 5 and 6. Female	
Gender parity index for gross enrolment ratio. Tertiary	Indicates the opportunity for females to enrol in higher education programs. Represents a ratio of GER-Female to GER-Male.
Distribution of students (%). ISCED 5A	Indicates how developed is a higher education sector in terms of the range of fields offered and the capacity of each field.
Distribution of students (%). ISCED 5B	
Distribution of students (%). ISCED 6	
Percentage of female students. ISCED 5A	Represents the number of female students enrolled in the specific ISCED level, expressed as a percentage of the total enrolment in the same level. Used to assess a gender disparity.
Percentage of female students. ISCED 5B	
Percentage of female students. ISCED 6	
Percentage of female students. Total	Represents the number of female students enrolled in higher education, expressed as a percentage of the total higher education level enrolment.
Gross completion rate. ISCED 5A. Total	Represents a number of graduates in ISCED 5A programs, expressed as a percentage of the population of the age at which students theoretically finish the most common ISCED 5A program in a country.
Gross completion rate. ISCED 5A. Male	
Gross completion rate. ISCED 5A. Female	
Gender parity index for gross completion rate. ISCED 5A	Indicates the opportunity for females to complete the ISCED 5A programs.
Percentage of female graduates in tertiary education	The number of female graduates, expressed as a percentage of all graduates in higher education.
[**IMR**] Inbound mobility rate	The number of students from abroad studying in a given country, expressed as a percentage of the total enrolment in higher education in that country.
[**OMR**] Outbound mobility ratio (%)	The number of students from a given country that study abroad as a percentage of the total enrolment in higher education in that country.
[**GOER**] Gross outbound enrolment ratio	The number of students from a given country that study abroad, expressed as a percentage of the higher education student age population in that country. Indicates a general level of participation in programs at foreign higher education institutions.

Attribute	Description
% of graduates in education	Distribution of graduates over different ISCED fields reflects the development of a higher education system in terms of the range of fields offered.
% of graduates in humanities and arts	
% of graduates in social sciences, business and law	
[**GRAD-SCI**] % of graduates in science	
% of graduates in engineering, manufacturing and construction	
% of graduates in agriculture	
% of graduates in health and welfare	
% of graduates in services	
% of graduates in unspecified programs	
Pupil-teacher ratio	Denotes an average number of students per professor, indicating the level of human resources dedicated to higher education.
[**TERT-STUD**] Tertiary students per 100,000 inhabitants	The number of students enrolled in higher education in a given academic-year, expressed per 100,000 inhabitants. The attribute indicates a density of students within the population of a country.
Female graduates as % of all graduates in education	Female graduates in each ISCED field as a percentage of all graduates in the same field.
Female graduates as % of all graduates in humanities and arts	
Female graduates as % of all graduates in social sciences, business and law	
Female graduates as % of all graduates in science	
Female graduates as % of all graduates in engineering, manufacturing and construction	
Female graduates as % of all graduates in agriculture	
Female graduates as % of all graduates in health and welfare	
Female graduates as % of all graduates in services	
Female graduates as % of all graduates in unspecified programs	

Expenditures – Higher Education	
Attribute	**Description**
Public expenditure per pupil as a % of GDP per capita. Tertiary	Public expenditure per student, expressed as a percentage of GDP per capita.
Educational expenditure by nature of spending as a % of total educational expenditure on public institutions. Tertiary. Salaries	The spending by nature, expressed as a percentage of the total expenditure on higher education. Salaries and other current add up to the total current expenditure.
Educational expenditure by nature of spending as a % of total educational expenditure on public institutions. Tertiary. Other current expenditure	
Educational expenditure by nature of spending as a % of total educational expenditure on public institutions. Tertiary. Total current expenditure	
Educational expenditure by nature of spending as a % of total educational expenditure on public institutions. Tertiary. Capital	
Total expenditure on educational institutions and administration as a % of GDP. All sources. Tertiary	Total expenditure on higher education, expressed as a percentage of GDP.
Total expenditure on educational institutions and administration as a % of GDP. Public sources. Tertiary	The spending on higher education, distributed by source, and expressed as a percentage of GDP.
Total expenditure on educational institutions and administration as a % of GDP. Private sources. Tertiary	
Percentage distribution of public current expenditure on education by level. Tertiary	Public current expenditure on higher education, expressed as a percentage of total public current expenditure on education. Indicates the relative emphasis of government spending on higher education within the overall educational expenditure.
Educational expenditure in tertiary as % of total educational expenditure	Expenditure for higher education, expressed as a percentage of total expenditure on education.

Expenditures – All Levels (Including Higher Education)	
Attribute	**Description**
[**PE-GNI**] Public expenditure on education as % of GNI	Proportion of country's wealth that has been spent on education during a given year.
[**PE-GDP**] Public expenditure on education as % of GDP	
[**PE-PUP-GDP-PC**] Public expenditure per pupil as a % of GDP per capita	Public expenditure per pupil/student, expressed as a percentage of GDP per capita.
Public expenditure on education as % of total government expenditure	Indicates government's policy emphasis on education relative to other public investments, showing how much the government invests into the development of human capital.
Public current expenditure on education as % of total current government expenditure	The share of total current government expenditure intended for current expenditure on education.
[**CE-GNI**] Current expenditure on education as % of GNI	Proportion of country's wealth that has been allocated for public current expenditures on education.
Public current expenditure on education as % of total public expenditure on education	Indicates the pattern of government spending on education in terms of the relative weight between the current and capital expenditures.
Total expenditure on educational institutions and administration as a % of GDP. All sources	Total expenditure on education, expressed as a percentage of GDP.
Total expenditure on educational institutions and administration as a % of GDP. Public sources	The spending on education, distributed by source, and expressed as a percentage of GDP.
Total expenditure on educational institutions and administration as a % of GDP. Private sources	
Total expenditure on educational institutions and administration as a % of GDP. International sources	

Constructed Attributes	
Attribute	**Description**
Theoretical/practical orientation of the majority of students	Constructed from the "Distribution of students (%)" – "ISCED 5A" and "ISCED 5B" attributes. If more students in a country study in the theoretically-oriented 5A programs than in the practically-oriented 5B programs, the attribute receives the value "theoretical". In the opposite situation, the attribute receives the value "practical".
GER-Total + GOER	A sum of the GER-Total and GOER attributes.
Popularity of a country for mobile students	Constructed from the IMR and OMR attributes. If more foreign students came to a given country to study than there are students from that country that left to study abroad, the attribute receives the value "popular", and vice versa (the value "unpopular").
Field of study completed by the most students	Constructed from the nine "Percentage of tertiary graduates" attributes by taking the name of a field containing those programs completed by the majority of students within a country. Accordingly, the attribute can take one of the nine values, e.g., education, humanities and arts, science.
Field of study completed by the least students	Constructed from the nine "Percentage of tertiary graduates" attributes by taking the name of a field containing those programs completed by the least of students within a country.
Expenditures – Higher Education	
The main source of investment in tertiary education	Constructed from the two attributes "Total expenditure on educational institutions and administration as a % of GDP. Tertiary" – "Public sources" and "Private sources". If the higher education sector is mostly financed from the public sources, the attribute received the value "public", and vice versa ("private").
Expenditures – All Levels (Including Higher Education)	
The level of education in which country invests the most	Constructed from the three attributes "Public expenditure per pupil as a % of GDP per capita" – Primary, Secondary and Tertiary. Indicates the level of education that is financially most supported by the country. Accordingly, it can take one of the three values: "primary", "secondary" and "tertiary".
Combined	
GER-Total + PE-GNI	A sum of GER-Total and PE-GNI attributes.
GER-Total + CE-GNI	A sum of GER-Total and CE-GNI attributes.

Appendix B: R&D Attributes

This appendix explains the R&D attributes used in the analyses presented in Section 5.2. The list of attributes is preceded by an explanation of general terms and categories that clarify the attributes. The explanations and definitions are extracted from the following sources: Frascati Manual (OECD, 2002), Manual for Statistics on Scientific and Technological Activities (UNESCO, 1984), UNESCO Institute for Statistics website (http://www.uis.unesco.org), World Bank website (http://data.worldbank.org) and WIPO website (http://www.wipo.int).

B.1 R&D Inputs

B.1.1 R&D Personnel

"R&D personnel" refers to the employees engaged in R&D activities, divided into:

- Researchers – Employees engaged in the conception or creation of new knowledge, products, processes, methods and systems, and in the management of projects.
- Technicians – Employees that perform scientific and technical tasks involving application of concepts and operational methods, normally under supervision of researchers.
- Other supporting staff – Skilled and unskilled craftsmen, secretarial and clerical staff associated with the R&D projects.

"R&D personnel" attributes are expressed in following quantities:

- FTE = Full time equivalent – Denotes person-years spent for R&D (e.g., a researcher employed 30% of his/hers total working hours on R&D activities represents 0.3 FTE). Attributes expressed in FTEs represent a true measure of the total volume of the R&D activities.
- HC = Head count – Represents the number of employees.

B.1.2 R&D Expenditures

Gross domestic expenditure on R&D (GERD) is total expenditure on R&D performed on the national territory during a year, including R&D performed within a country and funded from abroad, but excluding payments made abroad for R&D.
Purchasing Power Parities (PPP$) is a currency conversion method, which eliminates the differences in price levels among countries. GERD converted by the PPP method accounts for the same set of international prices of R&D activities; therefore, the comparisons between countries reflect only differences in the amount spent for R&D.

B.1.3. Institutional Classification Scheme

The "Institutional Classification Scheme" classifies R&D institutions into sectors based on the characteristic properties of their funding and activities (OECD, 2002). The sectors are:

- **Business enterprise** sector includes those institutions whose primary activity is the market production of goods and services, which are sold at an economically significant price.
- **Government** sector includes those institutions that: a) supply common services to the community (except higher education), which are not convenient to offer at market price; b) administer the state and the economic and social policy of the community.
- **Private non-profit** sector includes private non-market institutions serving to general public, as well as private individuals and households. Non-profit institutions provide services for the benefit of their members or for charity purposes, and are financed from membership subscriptions and donations.
- **Higher education** sector includes institutions of post-secondary education (universities, colleges of technology, etc.) no matter of their source of finance or legal status, as well as institutions such as research institutes, experimental stations and clinics associated with higher education institutions.
- **Abroad** includes institutions and individuals located outside of a country, as well as international organizations within a country that do not belong to business enterprise sector. It does not include vehicles, ships, aircrafts and space satellites that are abroad, but are operated by domestic entities.

B.2 R&D Outputs

An application for patent implies a submission of a form containing information about the applicant, the inventor and a specification of the form of intellectual property protection.
A grant of patent implies obtaining a set of exclusive rights when a patent is "granted".
High-technology export represents an export of products and services, which are the result of high intensity R&D activities, e.g., in aerospace industry, computer industry, pharmaceutical industry, in production of scientific instruments and electrical machinery.

B.3 Attributes

Attributes are divided into four groups that represent:

1. R&D personnel
2. R&D expenditures
3. R&D outputs
4. The previous three groups that are combined to form a set of constructed attributes

R&D Inputs – R&D Personnel	
Attribute	**Description**
Total R&D personnel (FTE)	The number of person-years spent for R&D activities.
R&D personnel – Female (FTE)	The number of person-years spent for R&D activities by females.
R&D personnel – Female (FTE) (%)	The percentage of person-years spent for R&D activities by females.
Total R&D personnel (HC)	The number of employees in the R&D sector.
R&D personnel – Female (HC)	The number of female employees in the R&D sector.
R&D personnel – Female (HC) (%)	The percentage of female employees in the R&D sector.
R&D personnel by sector of employment (FTE) – Business enterprise	The number of person-years spent for R&D activities by each of the four sectors.
R&D personnel by sector of employment (FTE) – Government	
R&D personnel by sector of employment (FTE) – Higher education	
R&D personnel by sector of employment (FTE) – Private non-profit	
Total researchers (FTE)	The number of person-years spent by the researchers for R&D activities.
Researchers – Female (FTE)	The number of person-years spent by the female researchers for R&D activities.
Researchers – Female (FTE) (%)	The percentage of person-years spent by the female researchers for R&D activities.
[**RES-FTE**] Researchers per million inhabitants (FTE)	Conveys the same information as "Total researchers (FTE)" attribute, but this time expressed per million inhabitants.
Total researchers (HC)	The number of researchers in the R&D sector.
Researchers – Female (HC)	The number of female researchers in the R&D sector.
Researchers – Female (HC) (%)	The percentage of female researchers in the R&D sector.
[**RES-HC**] Researchers per million inhabitants (HC)	Conveys the same information as "Total researchers (HC)" attribute, but this time expressed per million inhabitants.

Attribute	Description
Researchers by sector of employment (FTE) – Business enterprise	The number of person-years spent by the researchers for R&D activities, within each of the four sectors.
Researchers by sector of employment (FTE) – Government	
Researchers by sector of employment (FTE) – Higher education	
Researchers by sector of employment (FTE) – Private non-profit	
Total technicians (FTE)	The number of person-years spent by the technicians for R&D activities.
Technicians – Female (FTE)	The number of person-years spent by the female technicians for R&D activities.
Technicians per million inhabitants (FTE)	Conveys the same information as "Total technicians (FTE)", but this time expressed per million inhabitants.
Total technicians (HC)	The number of technicians in the R&D sector.
Technicians – Female (HC)	The number of female technicians in the R&D sector.
Technicians per million inhabitants (HC)	Convey the same information as "Total technicians (HC)", but this time expressed per million inhabitants.
Total other supporting staff (FTE)	The number of person-years spent by the supporting staff for R&D activities.
Other supporting staff – Female (FTE)	The number of person-years spent by the female supporting staff for R&D activities.
Total other supporting staff (HC)	The number of "supporting staff" employees in the R&D sector.
Other supporting staff – Female (HC)	The number of female "supporting staff" employees in the R&D sector.

R&D Inputs – R&D Expenditures	
Attribute	**Description**
[**GERD-PC**] GERD per capita (PPP$)	GERD expressed in different quantities.
[**GERD-GDP**] GERD as % of GDP	
Total GERD (000 PPP$)	
[**SRC-BE**] Source of funds for R&D – Business enterprise (%)	The distribution of GERD over the five sectors.
Source of funds for R&D – Government (%)	
Source of funds for R&D – Higher education (%)	
Source of funds for R&D – Private non-profit (%)	
Source of funds for R&D – Funds from abroad (%)	
Source of funds for R&D – Not distributed funds (%)	
R&D Outputs	
Applications for patents (residents)	The number of applications for patents where the first applicant is a resident of a country.
[**APP-NON-RES**] Applications for patents (non-residents)	The number of applications for patents where the first applicant is a non-resident of a country.
Applications for patents (total)	The total number of applications for patents submitted by a country.
Grants of patents (residents)	The number of granted patents where the first applicant is a resident of a country.
Grants of patents (non-residents)	The number of granted patents where the first applicant is a non-resident of a country.
Grants of patents (total)	The total number of granted patents for a country.
[**HI-TECH**] High-technology exports (% of manufactured exports)	The percentage of high-technology exports in total exports of a country.

Constructed Attributes	
R&D Inputs – R&D Personnel	
Attribute	**Description**
[**R&D-PERS-FTE**] R&D personnel per million inhabitants (FTE)	Constructed by dividing the "Total R&D personnel (FTE)" with a population of a country and by multiplying the result with a million.
[**R&D-PERS-HC**] R&D personnel per million inhabitants (HC)	Constructed by dividing the "Total R&D personnel (HC)" with a population of a country and by multiplying the result with a million.
[**SEC-R&D-PERS**] Sector employing the most R&D personnel	Indicates which sector performs the most R&D activities in a country. Constructed from the four "R&D personnel by sector of employment (FTE)" attributes. Considering that "private non-profit" sector never appears as the sector that employs the most R&D personnel, the newly constructed attribute can take one of the four values: business enterprise, government, higher education and N/A (not known).
[**SEC-RES**] Sector employing the most researchers	Similar to the SEC-R&D-PERS attribute, except that it accounts only for work of researchers, and not of technicians and other supporting staff.
R&D Inputs – R&D Expenditures	
[**GERD-GNI**] GERD as % of GNI	Constructed using the following formula: "Total GERD (000 PPP$)" * 1000) / "GNI per capita (PPP$)".
[**SEC-INVEST**] Sector investing the most in R&D	Indicates which sector invests the most in R&D activities. Constructed from the six "Source of Funds for R&D" attributes. Accordingly, it can take one of the six values: business enterprise, government, higher education, private non-profit, abroad and N/A (not known or not distributed funds).
R&D Outputs	
[**APP-RES-%**] Applications for patents (residents) (%)	Constructed by dividing "Applications for patents (residents)" with "Applications for patents (total)".
[**APP-NON-RES-%**] Applications for patents (non-residents) (%)	Constructed by dividing "Applications for patents (non-residents)" with "Applications for patents (total)".

Attribute	Description
Majority of applications for patents (residents - non-residents)	Indicates whether the residents or non-residents of a country submit the most applications for patents, under the assumption that the first author represents the leader in a patent creation process. The attribute can take one of the tree values: residents, non-residents and N/A.
Applications for patents per researcher (FTE)	Constructed by dividing "Applications for patents (total)" with "Total researchers (FTE)".
[**APP-HC**] Applications for patents per researcher (HC)	Constructed by dividing "Applications for patents (total)" with "Total researchers (HC)".
[**APP-PMI**] Applications for patents per million inhabitants	Constructed by dividing "Applications for patents (total)" with population of a country and by multiplying the result by a million.
Grants of patents (residents) (%)	Constructed by dividing "Grants of patents (residents)" with "Grants of patents (total)".
Grants of patents (non-residents) (%)	Constructed by dividing "Grants of patents (non-residents)" with "Grants of patents (total)".
Majority of grants of patents (residents - non-residents)	Indicates whether the residents or non-residents of a country have more granted patents, under the assumption that the first author represents the leader in a patent creation process. The attribute can take one of the tree values: residents, non-residents and N/A.
Grants of patents per researcher (FTE)	Constructed by dividing "Grants of patents (total)" with "Total researchers (FTE)".
Grants of patents per researcher (HC)	Constructed by dividing "Grants of patents (total)" with "Total researchers (HC)".
[**GRANT-PMI**] Grants of patents per million inhabitants	Constructed by dividing "Grants of patents (total)" with population of a country and by multiplying the result by a million.
Grants of patents per application for patent	Constructed by dividing "Grants of patents (total)" with "Applications for patents (total)".

Appendix C: Questionnaire

General instructions: Please fill one page after another without looking ahead. Looking backwards is allowed, but not modifying your answers.

The data consists of 167 learning examples, each representing a description of a country and a class (GNI per capita – low, middle, high). There are 27 attributes describing the R&D sector. The motivation is to find which attributes and relations contribute the most to the economic welfare of a country.

The first tree in Figure C.1 was constructed in Weka. Please take a look at the tree, examine all the nodes in the tree and reply to the questions below. In the leaf there are two numbers. The first number denotes the number of examples in the leaf, and the second number represents the number of examples of non-majority classes.

While most of the attributes are comprehensible, GERD needs a short explanation: "GERD per capita" represents the level of investment in R&D in the relative form (PPP$ – purchasing power parity) to avoid direct link to the economic welfare, and "GERD as % of GNI" denotes percentage of GNI (gross national income), designated to research.

CCPE denotes corrected class probability estimate and is a measure showing how significant the tree is in comparison to all possible trees constructed from this data. 1 is the most significant and 0 the least. Accuracy and Kappa are measured in a 10-fold cross-validation.

Questions:

1. Does the tree sound reasonable or not?

   YES NO

2. Is the attribute in the root node (that the most important factor for the welfare of a country is the level of investment in R&D) reasonable or not?

   YES NO

3. Does the right subtree (starting with "Sector investing the most in R&D") present reasonable relations or not?

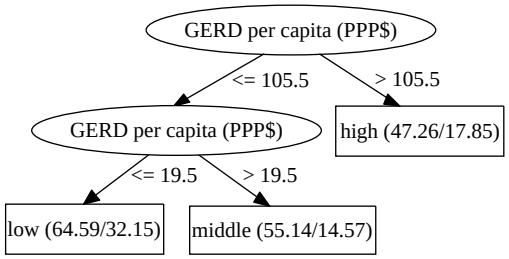
   YES NO

4. Does the left subtree (starting with "Sector employing the most researchers") present reasonable relations or not?

   YES NO

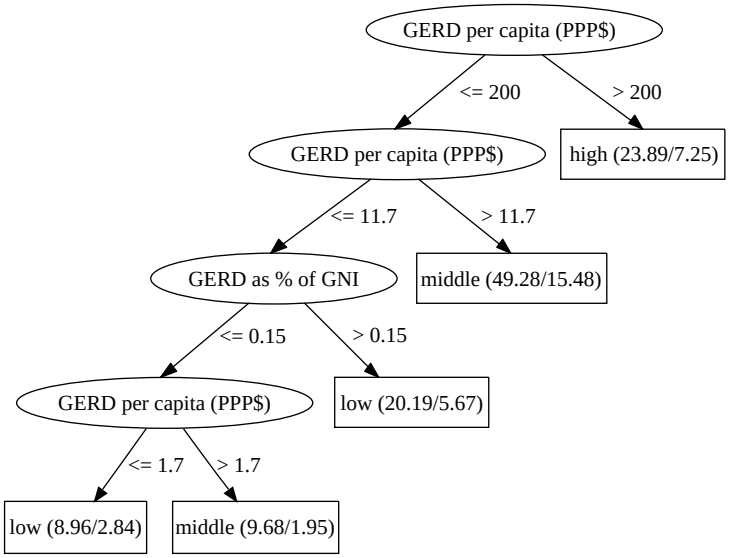
Figure C.1: The first tree.

Three additional trees were generated with accuracy 63.47% (−1.2 percentage points compared to the tree on the previous page) and the other two with accuracy 66.47% (+1.8). Please take a look at the trees, examine all the nodes in the trees and reply to the questions on the next page. The questions from the first page are repeated, but this time additional information is provided.



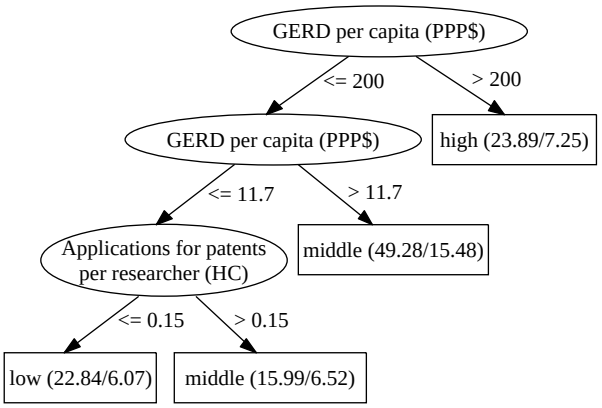
Accuracy : 63.47% (-1.2)
Kappa: 0.4023 (-0.0024)
Complexity : 3 relations (-8)
CCPE: 0.5685 (+0.1572)

Figure C.2: The second tree.



Accuracy : 66.47% (+1.8)
Kappa: 0.4314 (+0.0267)
Complexity : 5 relations (-6)
CCPE: 0.5287 (+0.1174)

Figure C.3: The third tree.



Accuracy : 66.47% (+1.8) Complexity : 4 relations (-7)
Kappa: 0.432 (+0.0273) CCPE: 0.5175 (+0.1062)

Figure C.4: The fourth tree.

Questions:

5. Which of the additional three trees sounds more reasonable than the first one? *Please select one or more answers (e.g., c and d means that the third and the fourth tree sound more reasonable than the first one).*
   a. none
   b. second
   c. third
   d. fourth

6. What is your decision based upon? *Please select one answer.*
   a. accuracy
   b. complexity
   c. other measures: CCPE and/or Kappa
   d. content of the tree
   e. a combination of the above (e.g., b, c):_____

7. Write down the sequence of trees that persuaded you the most that the GERD attributes are the most important for the welfare of a country? E.g., 2, 3, 1, 4 means that 2 was the most persuasive and 4 the least.

   _____

8. Additional information to question 3 from the first page: Five leaves in the right-hand side of the tree contain class "high" and only one "middle". Analysis of the learning data reveals that there is actually no example with class "middle" in the data set corresponding to this subtree, but a couple of unknown values which the algorithm distributes among the default values cause this effect. Therefore, the whole right subtree could be represented by a leaf "high". In light of this information please reply to the two questions.

   Do you agree with the explanation above?

   YES NO

   Does the right subtree (Sector investing the most in R&D) present a trust-worthy relation or not?

   YES NO

9. Additional information to question 4 from the first page: Four leaves in the left-hand side of the tree contain class "middle" and only one "low". Would not it be reasonable to substitute the node "Sector employing the most researchers" with a node below (GERD per capita), as in the second tree? In light of this information please reply to the next questions.

   Do you agree with the explanation above?

   YES NO

   Does the left subtree (Sector employing the most researchers) present trust-worthy relations or not?

   YES NO

   Would you prefer to substitute the left-hand subtree of the first tree with any of the left-hand subtrees of the third or the fourth tree?

   YES NO

   If the answer to the previous question was positive, please answer from which tree would you prefer to use the subtree? Please select one answer.

   THIRD FOURTH ANY

Appendix D: Description of Interactions with the System Implementing the HMDM method

The HMDM method is embedded in an interactive system coded in Java. In the following paragraphs, we will present the elements of a graphical user interface (Section D.1) and an example of a typical interaction between the user and the system (Section D.2).

D.1 Interface

The interface is composed of a main widow with a menu and a toolbox. The main widow is populated with task-specific windows chosen by the user.

Within the system, a domain analysis is represented with a project. Projects are managed from the "Project" menu (see Figure D.1). Each project defines a path to a data set in ARFF format (Witten and Frank, 2005) and a folder for storing results of the analysis. By storing the already computed results, the user may choose to continue the analysis any time without loosing already constructed models.

Figure D.1: The menu and the toolbox.

The analysis is initiated by clicking on the blue arrow button in the toolbox (see Figure D.1) that leads to a window in Figure D.2, which represents the initial DM step. At the top of the window, the user first defines the DM method, parameters and their ranges, and constraints. By clicking on the button "Start Initial DM", all the possible models are constructed with the selected DM method and the defined parameters that satisfy the constraints. A summary of the constructed models, ordered by the predefined criteria, are presented in the table. For each model, the table presents the parameters used to construct the model, quality (ACC, Kappa, CPE and CCPE) and complexity (CPX) of the model, as well as the list of attributes from the model. By clicking on a model in the table, a preview of the model appears right of the table. The user can select one or several of the models for further analysis, which are then added to the list below the table. By clicking on the button "Analyze selected models", for each of the selected models, a separate relation-analysis session is created.

Within the relation-analysis session, the selected model takes the role of the initial model and serves as a starting point in the process of searching for credible relations. The initial model is presented in the window in Figure D.3, together with removed and added attributes graphs. The graphs are used to modify the initial model in order to design relations and examine their credibility. The graphs are populated either automatically, or interactively. Automatic population is executed by clicking on the "Remove attributes" and "Add attributes" buttons, which are located right of the initial model. The two buttons call the procedures REMOVE_ATTRIBUTES and ADD_ATTRIBUTES. The interactive population is explained later in the text.

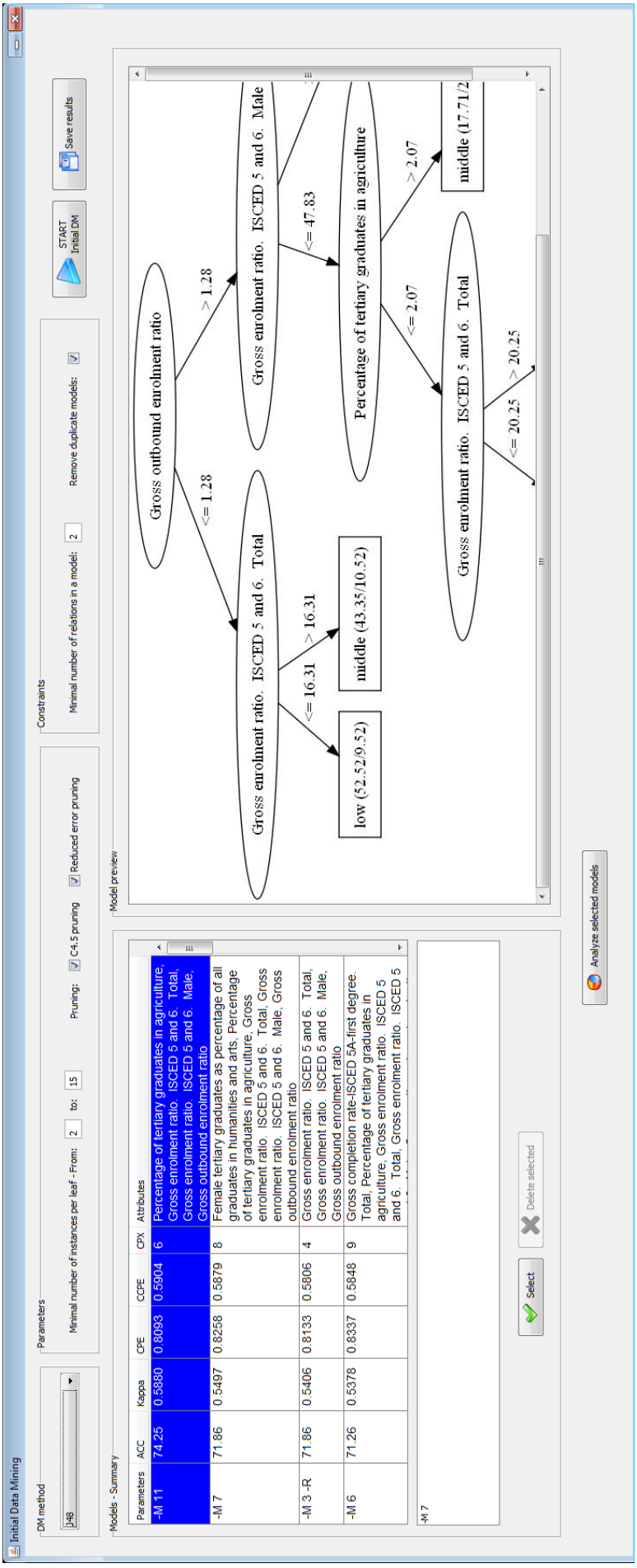


Figure D.2: The window that represents Initial DM.

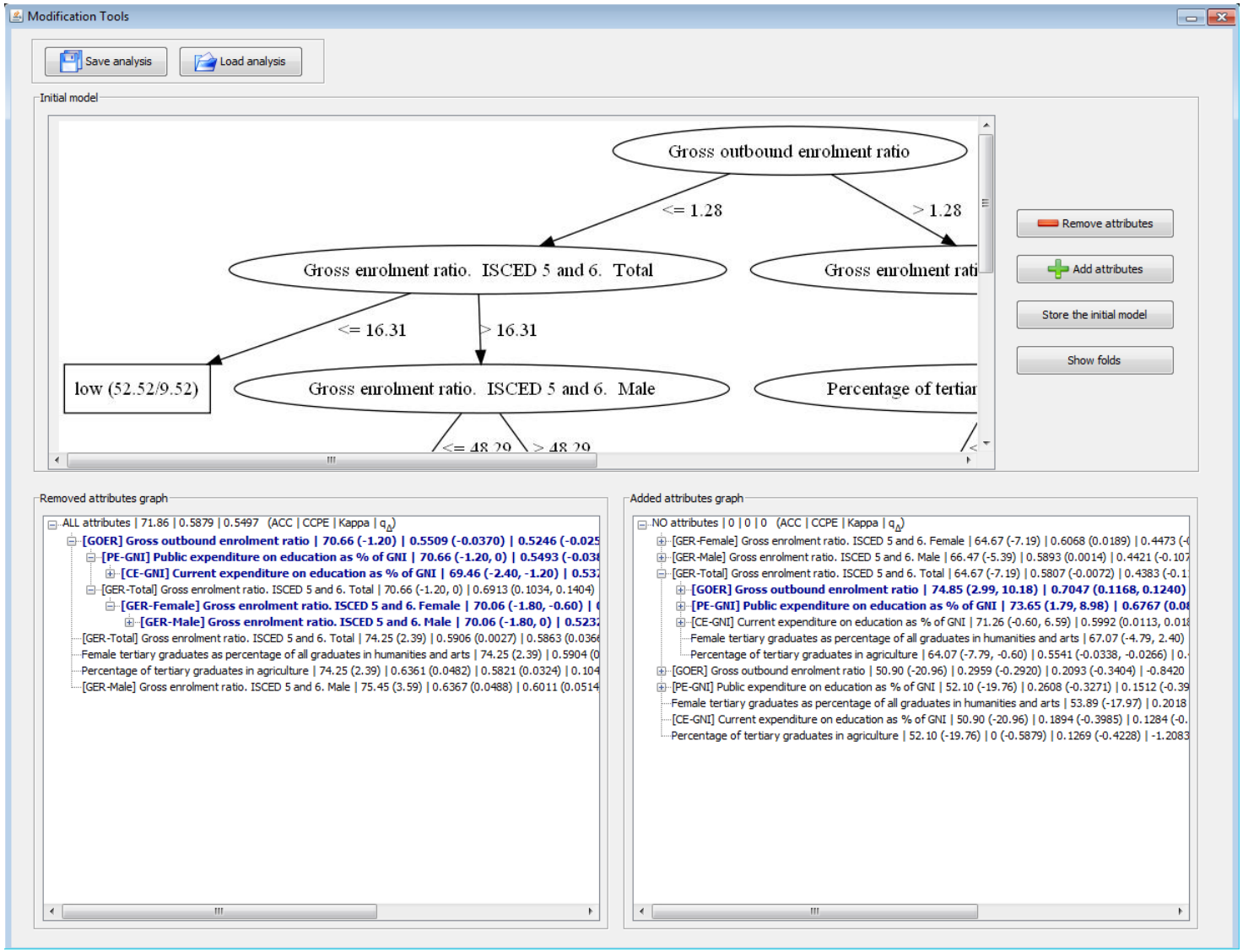


Figure D.3: The window that represents analysis of the relations.

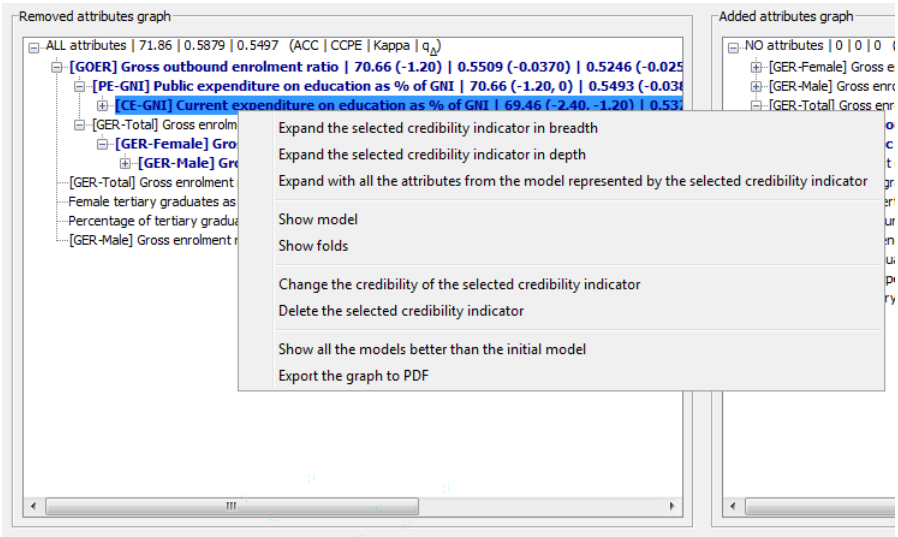


Figure D.4: An example of a remove attributes graph with active context menu.

Right of the initial model, there are two tools contextually connected with the initial model: "Store the initial model" and "Show folds". When the user confirms the credibility of relations within the initial model, he/she stores the model by clicking the button "Store the initial model". By clicking on the "Show folds" button, the user can examine a set of models constructed for cross-validation folds. This tool clarifies the role of attributes that appear in the initial model, but do not modify the quality when removed. In most of the cases such attributes do not even appear within the models constructed for different folds, which indicates less-credible attributes. In some cases, there may be some other causes to this phenomenon and this tool helps to clarify such situations.

The removed and added attributes graphs are interactive explanations that contain tools for designing and re-examining relations for credibility. The tools are available from a context menu, which is obtained by right-clicking the credibility indicator (a node in the graph). An example of the removed attributes graph with active context menu is presented in Figure D.4. The tools are divided into four groups that represent different functions. The first group contains three "expand credibility indicator" tools:

- Expand the selected credibility indicator in breadth – This tool opens a new window where the user selects one or several of the attributes from the data set. Then, the selected attributes are added to those level of the graph that is subordinated to the selected credibility indicator.

- Expand the selected credibility indicator in depth – Similar to the previous tool, the user first selects the attributes, but this time, he/she also defines the order in which the attributes are added to the graph. The attributes are then added in depth, following the user-defined order.

- Expand with all the attributes from the model represented by the selected credibility indicator – This tool is applicable only to the removed attributes graph. The attributes from the model, represented by the selected credibility indicator, are extracted and added to the level of the graph subordinated to the selected indicator.

The second group contains two tools for viewing models represented by the indicator:

- Show model – Presents the model represented by the selected credibility indicator.

- Show folds – Presents the models constructed for cross-validation folds. In contrast to the tool represented by the "Show folds" button, this tool constructs per-fold models from a combination of attributes specified by the selected indicator.

The third group contains the tools for managing the credibility of the selected indicator:

- Change the credibility of the selected credibility indicator – A credibility of an indicator is determined by applying the rules implemented within the system. However, the rules do not formalize criteria such as common sense. By examining a model represented by the indicators, the user establishes the meaning of relations within and changes the credibility status when the system made a wrong suggestion.
- Delete the selected credibility indicator – In case of large graphs, the user may prefer to delete less-credible indicators from the graph. With this tool he/she can delete any indicator, together with all of the subordinate indicators.

The fourth group contains two tools that influence all the indicators in the graph:

- Show all the models better than the initial model – Presents all the models for which the q_Δ measure is positive.
- Export the graph to PDF – The user can export the constructed graph to a PDF file.

D.2 An Example of a Typical Interaction

Suppose that the user selected the data and the initial model presented in Subsection 5.1.1. The typical interaction with the presented interactive system would include the following steps:

1. Select the automatic procedure for removing attributes by clicking on the "Remove attributes" button.
2. Examine the constructed indicators within the removed attributes graph for credibility – quality-based comparisons are made with the help of the quality measures and differences in quality presented within the indicators, while meaning is established by observing the models represented by the indicators ("Show model" tool). Change the credibility status of those indicators, which were mistakenly marked by the system as such ("Change the credibility of the selected credibility indicator" tool).
3. Iteratively select any of the tools from the "expand credibility indicator" group to examine the credibility of those attribute combinations that were not constructed by the automatic method. Re-establish the credibility of the constructed indicators based on a new evidence.
4. With the help of the removed attributes graph, the user made hypotheses about credible relations in the domain. Attributes from those relations are typically selected as an input into the ADD_ATTRIBUTES procedure in order to obtain an additional evidence that will confirm or reject the relations. The input is defined through the menu obtained by clicking on the button "Add attributes". The other two choices for the input are the set of attributes from the initial model and an arbitrary set of attributes.

5. The user repeats the second and the third step, but this time on the added attributes graph. When a new interesting relation emerges within the added attributes graph the user may re-examine its credibility with the help of the removed attributes graph.

6. When no more interesting relations are found near the initial model, the user concludes the analysis for the selected initial model. The constructed graphs can be stored by clicking on the "Save analysis" button and reviewed later on by clicking on the "Load analysis" button (the top of the Figure D.3). The graphs can be further reduced by deleting less-credible indicators (using "Delete the selected credibility indicator" tool) and exported to a PDF file.

In addition, at any point during the analysis, the user can review the list of stored credible models by clicking on the second button in Figure D.1, and revise the list in the light of a new evidence.

Appendix E: Bibliography

E.1 Publications Related to This Thesis

E.1.1 Journal papers:

- [Vidulin and Gams (2011)] Vidulin, V.; Gams, M. Impact of High-Level Knowledge on Economic Welfare through Interactive Data Mining. *Applied Artificial Intelligence* **25**(4):267–291 (2011).

- [Vidulin et al. (2009)] Vidulin, V.; Luštrek, M.; Gams, M. Multi-Label Approaches to Web Genre Identification. *Journal for Language Technology and Computational Linguistics* **24**(1):93–110 (2009).

- [Vidulin et al. (2007b)] Vidulin, V.; Luštrek, M.; Gams, M. Training a Genre Classifier for Automatic Classification of Web Pages. *Journal of Computing and Information Technology* **15**(4):305–311 (2007b).

- [Vidulin and Gams (2006a)] Vidulin, V.; Gams, M. Analyzing the Impact of Investment in Education and R&D on Economic Welfare with Data Mining. *Electrotechnical Review* **73**(5):285–290 (2006a).

E.1.2 Conference papers:

- [Vidulin and Gams (2010)] Vidulin, V.; Gams, M. Searching for Meaningful Models in Macroeconomic Domain. In: *Proceedings of the 13th International Multiconference Information Society – IS 2010*, 94–97 (2010).

- [Vidulin and Gams (2009)] Vidulin, V.; Gams, M. Multi-Label Classification of Web Genres. In: *Proceedings of the 18th International Electrotechnical and Computer Science Conference – ERK 2009*, 179–182 (2009).

- [Vidulin (2009)] Vidulin, V. Problem Transformation Methods for Multi-Genre Web Pages Classification. In: *Proceedings of the 12th International Multiconference Information Society – IS 2009*, 136–139 (2009).

- [Vidulin and Gams (2008a)] Vidulin, V.; Gams, M. Impacts of Education and R&D on Economy: Analysis by Data Mining Techniques. In: *Proceedings of the International Conference on Advances in the Internet, Processing, Systems, and Interdisciplinary Research – VIPSI 2008* (2008a).

- [Vidulin and Gams (2008b)] Vidulin, V.; Gams, M. Is Science Important for Economic Welfare? In: *Proceedings of the 11th International Multiconference Information Society – IS 2008*, 41–44 (2008b).

- [Rehm et al. (2008)] Rehm, G.; Santini, M.; Mehler, A.; Braslavski, P.; Gleim, R.; Stubbe, A.; Symonenko, S.; Tavosanis, M.; Vidulin, V. Towards a Reference Corpus

of Web Genres for the Evaluation of Genre Identification Systems. In: *Proceedings of the 6th International Language Resources and Evaluation – LREC 2008*, 351–358 (2008).

- [Vidulin and Gams (2007)] Vidulin, V.; Gams, M. The Impact of High Level Knowledge on Economic Welfare. In: *Proceedings of the 10th International Multiconference Information Society – IS 2007*, 107–110 (2007).

- [Luštrek et al. (2007)] Luštrek, M.; Vidulin, V.; Gams, M. Genres for Web Page Classification. In: *Proceedings of the International Conference on Advances in the Internet, Processing, Systems, and Interdisciplinary Research – VIPSI 2007* (2007).

- [Vidulin et al. (2007a)] Vidulin, V.; Luštrek, M.; Gams, M. Evaluation of Different Approaches to Training a Genre Classifier. In: *Proceedings of the 2007 International Conference on Artificial Intelligence and Pattern Recognition – AIPR 2007*, 515–520 (2007a).

- [Vidulin et al. (2007c)] Vidulin, V.; Luštrek, M.; Gams, M. Training the Genre Classifier for Automatic Classification of Web Pages. In: *Proceedings of the 29th International Conference on Information Technology Interfaces – ITI 2007*, 93–98 (2007c).

- [Vidulin et al. (2007d)] Vidulin, V.; Luštrek, M.; Gams, M. Using Genres to Improve Search Engines. In: *Towards Genre-Enable Search Engines: the Impact of Natural Language Processing – Proceedings of the International Workshop*, 45–51 (2007d).

- [Vidulin and Gams (2006b)] Vidulin, V.; Gams, M. Impact of Investment in Education and R&D on Economic Growth. In: *Proceedings of the 15th International Electrotechnical and Computer Science Conference – ERK 2006*, 129–132 (2006b).

- [Vidulin et al. (2006)] Vidulin, V.; Luštrek, M.; Gams, M. Comparison of the Performance of Genre Classifiers Trained by Different Machine Learning Algorithms. In: *Proceedings of the 9th International Multiconference Information Society – IS 2006*, 140–143 (2006).

E.1.3 Popular articles:

- [Gams and Vidulin, 2006] Gams, M.; Vidulin, V. Vpliv znanja na gospodarsko uspešnost. *Finance* 2006/166, 18 (30.8.2006).

E.2 Other Publications

E.2.1 Conference papers:

- [Vidulin and Filipič (2006)] Vidulin, V.; Filipič, B. Visualization of a Simple Genetic Algorithm for Pedagogical Purposes. In: *Proceedings of the 15th International Electrotechnical and Computer Science Conference – ERK 2006*, 99–102 (2006).

E.2.2 Professional articles:

- [Vidulin (2006)] Vidulin, V. Constructivist Learning Theory as a Link Between Artificial Neural Networks and Intelligent Tutoring Systems. *Organizacija* **39**(2):154–156 (2006).

Appendix F: Biography

Vedrana Vidulin was born in Rijeka, Croatia, on April 12, 1981. She received a university degree in 2005 from the Faculty of Philosophy, University of Rijeka, Croatia, by defending the thesis "Neural Networks: Algorithms and Applications in Education". She was awarded with the Rector's Award of the University of Rijeka as the best student of senior years enrolled in programs of the Faculty of Philosophy.

In 2005, she enrolled in the "New Media and E-Science" program at the Jožef Stefan International Postgraduate School, Ljubljana, Slovenia. She holds a scholarship for doctoral studies awarded by the Slovene Human Resources Development and Scholarship Fund and a scholarship of Department of Intelligent Systems, Jožef Stefan Institute, Ljubljana, Slovenia.

Since December 2005, she has worked at the Jožef Stefan Institute in the Department of Intelligent Systems, under the supervision of prof. dr. Matjaž Gams. Her research focuses on the development of novel machine learning algorithms, with the emphasis on two domains: analysis of macroeconomic data and text categorization. In the text categorization field, the initial emphasis was on the classification of web pages with genres. The developed genre classificator was implemented within the ALVIS, a semantic search engine (FP6 European project). SEMEA – Semantic Multi-Label Evolutionary Algorithm is the algorithm conceived for the task of web genre classification, which has evolved into general-purpose text categorization algorithm. The paper on this work was published in international "Journal for Language Technology and Computational Linguistics". The results of the analysis of macroeconomic data, together with the Human-Machine Data Mining algorithm, were published in the SCI journal "Applied Artificial Intelligence". The paper presenting the improvements of the algorithm is submitted to the SCI journal "International Journal of Information Technology and Decision Making".