Information Sciences 288 (2014) 254–278

Contents lists available at ScienceDirect

Information Sciences

journal homepage: www.elsevier.com/locate/ins

Combining human analysis and machine data mining to obtain credible data relations

Vedrana Vidulin *, Marko Bohanec, Matjaž Gams

Jožef Stefan Institute, Jamova cesta 39, SI-1000 Ljubljana, Slovenia

ARTICLE INFO

Article history:
Received 19 November 2012
Received in revised form 25 July 2014
Accepted 2 August 2014
Available online 12 August 2014

Keywords:
Interactive data mining
Interactive machine learning
Interactive explanation structure
Relation-extraction scheme
Domain analysis
Human–computer interaction

ABSTRACT

Can a model constructed using data mining (DM) programs be trusted? It is known that a decision-tree model can contain relations that are statistically significant, but, in reality, meaningless to a human. When the task is domain analysis, meaningless relations are problematic, since they can lead to wrong conclusions and can consequently undermine a human's trust in DM programs. To eliminate problematic relations from the conclusions of analysis, we propose an interactive method called Human–Machine Data Mining (HMDM). The method constructs multiple models in a specific way so that a human can reexamine the relations in different contexts and, based on observed evidence, conclude which relations and models are credible—that is, both meaningful and of high quality. Based on the extracted credible relations and models, the human can construct correct overall conclusions about the domain. The method is demonstrated in two complex domains, extracting credible relations and models that indicate the segments of the higher education sector and the research and development sector that influence the economic welfare of a country. An experimental evaluation shows that the method is capable of finding important relations and models that are better in both meaning and quality than those constructed solely by the DM programs.

1. Introduction

In data mining (DM) and machine learning (ML), a human supplies the data and tunes the parameters of used methods. The obtained model is typically the result of several iterative, parameter-tuning steps. This paper aims to improve the interaction between humans and DM and ML programs and, therefore, belongs to the field of interactive DM (IDM) or interactive ML (IML) (the terms are used interchangeably in the literature) [71,73].

The goal of IML is to "help scientists and engineers exploit more of their specialized data" [53]. IML "focuses on methods that empower domain experts to control and direct machine learning tools from within the deployed environment, whereas traditional machine learning does this in the development environment" [53].

The field of IML has recently received a great deal of attention. The preface of the IUI 2013 Workshop on Interactive Machine Learning stated, "Many applications of Machine Learning (ML) involve interactions with humans... a growing community of researchers at the intersection of ML and human–computer interaction are making interaction with humans a central part of developing ML systems. These efforts include applying interaction design principles to ML systems, using

* Corresponding author. Tel.: +386 1 477 3147; fax: +386 1 477 3131.
E-mail addresses: vedrana.vidulin@ijs.si (V. Vidulin), marko.bohanec@ijs.si (M. Bohanec), matjaz.gams@ijs.si (M. Gams).

http://dx.doi.org/10.1016/j.ins.2014.08.014
0020-0255/

human-subject testing to evaluate ML systems and inspire new methods, and changing the input and output channels of ML systems to better leverage human capabilities" [54]. The mission statement of one of the Microsoft research groups dealing with IML notes: "... with the advancement of computational techniques such as machine learning, we now have the unprecedented ability to embed 'smarts' that allow machines to assist users in completing their tasks. We believe that trying to fully automate tasks is extremely difficult and even undesirable, but instead there exists a computational design methodology which allows us to gracefully combine automated services with direct user manipulation" [12].

When supervised DM and ML methods construct models in complex domains, such as economic and social domains, the models often contain *less-credible relations* [1,27,57]. Here, the *relation* is a pattern that connects a set of attributes describing the properties of a concept underlying the data with a class/target attribute, which represents the concept. The term *less-credible* means that the relation is of either low quality or high quality, but is meaningless to a human analyst. *Meaningless* means that a relation's semantic is contradictory to the human's common sense or domain knowledge, and a meaningless state can only be determined by including the human in the DM process. When the task is domain analysis, less-credible relations must be eliminated from the constructed models, since they lead to incorrect conclusions about the most important relations in the domain and can, consequently, undermine the human's trust in the DM system [59].

The problem is illustrated by the example in Fig. 1. The decision-tree model on the right side of Fig. 1 represents a domain model. The tree is constructed from the data (the table on the left side of Fig. 1) using the J48 algorithm in Weka [69] with default parameters. The first three columns, or attributes, of this table represent properties of a person, while the final column, or class, indicates the person's gender. Each row, or example, represents a person. In the tree, the node represents an attribute, and the leaves represent the class. In each leaf, the number in brackets represents the number of examples that reach that leaf. The tree contains a single relation, indicating that a person is a woman if the person has long hair and that a person is a man if the person has short hair. The relation is of high quality, since the tree's accuracy (ACC) is 100%. ACC denotes the overall performance of the tree, expressed as the percentage of correctly classified examples out of all the examples classified by the tree. The relation is meaningless, however, since several men have long hair but are not women (as the left branch of the tree suggests).

The problem that this paper examines most commonly stems from an incompleteness of data [52]. For example, adding more rows and columns to the table in Fig. 1 would likely result in a different relation, but adding the right additional data might be a demanding task. Humans, however, can detect weak relations in domain models using domain knowledge and common sense.

The knowledge that men and women have long and short hair is objective in terms of common sense, as is the case in Fig. 1, but it is hard to take a purely objective position when humans are involved. Humans can also be subjective in terms of fairness; however this discussion is beyond the scope of this paper.

Although the relation in Fig. 1 is of high quality, its meaninglessness makes it less credible.

Another example was obtained through DM in a real-life domain. The decision-tree model presented in Fig. 2 is constructed with the J48 algorithm in Weka using the default parameters and a minimum number of instances per leaf (MNIL) of 5. The tree is constructed from a data set composed of 37 attributes describing the research and development (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" (see Section 4.1 for more information on this data set). In the tree, the subtrees form the relations. In 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 of the class value other than the one represented by the leaf. The quantities are expressed in decimals to account for the weights of the examples with missing values.

The tree contains three interesting relations. The first is that countries with better welfare invest extensively in R&D. The relation contains attribute "GERD per capita (PPP$)" (GERD stands for Gross Domestic Expenditure on R&D and PPP$ for purchasing power parity in American dollars), which represents the level of investment in R&D. This relation appears twice in the tree. Both times, the "higher than" side of the subtree (>10.8 and >105.5) leads to leaves representing welfare better than that on the "less than" side. One could conclude that the first relation is a valid candidate for a *credible relation* in the tree because it is *meaningful*; that is, it is in accordance with domain knowledge [63] and common sense, it appears twice in the tree and, both times, it makes a clear distinction between countries with different levels of welfare. This relation is marked in bold in Fig. 2. The second relation—"Sector investing the most in R&D" (the right subtree)—seems to be meaningless, since all but one of the leaves represent the class "high", and the single "middle" leaf represents the countries for which the sector is unknown ("N/A" value). Therefore, the entire subtree can be replaced with a single node: "high". A detailed analysis shows

Wears a skirt?	Has long hair?	Has a woman-like body shape?	Sex of a person?
yes	yes	yes	woman
no	yes	no	woman
yes	yes	yes	woman
yes	yes	no	woman
no	no	no	man
no	no	no	man

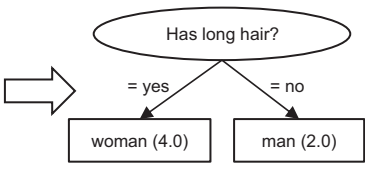

Fig. 1. An example of a domain model ("women have long hair and men do not"), correctly constructed from incomplete data.

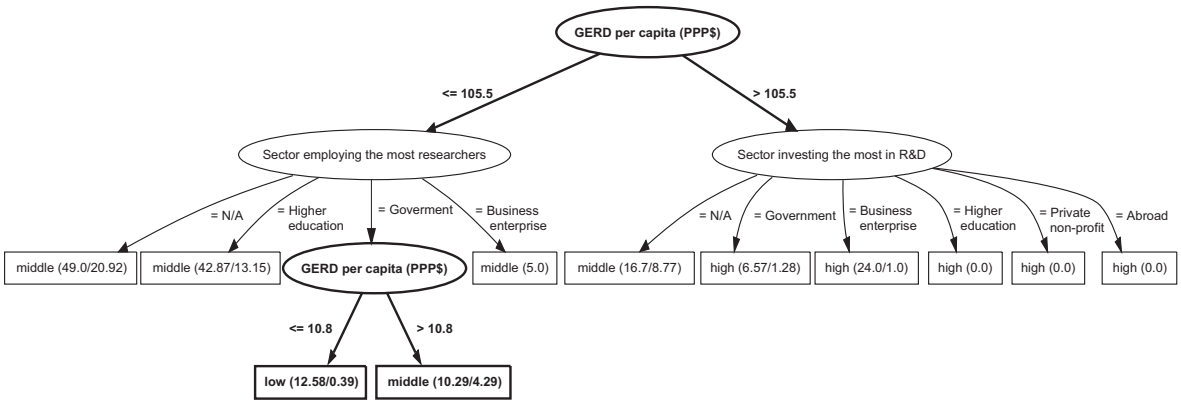

Fig. 2. Decision-tree model constructed for an R&D domain.

that the problem is caused by several missing values, resulting in a relation that is statistically correct but meaningless to humans. The third relation—"Sector employing the most researchers"—distinguishes between "low" and "middle" countries. However, for "middle" countries, any sector could be the main employer, which makes the relation meaningless. Due to their lack of meaning, the second and third relations are not shown in bold. However, these relations should be verified with additional tests.

To eliminate less-credible relations from the models, both automatic and interactive approaches were suggested. Examples of the former include the pruning of decision trees [56], maximum-ambiguity-based sample selection for tree construction [67], alternative-node-selection measures in trees [70], fuzzy-entropy-maximization-based classification rule refinement [66], and the correction of a quality estimate to eliminate the random rules with optimistically high values of quality [47]. Typical examples of the latter suggest improvements in the form of new training examples [18] or a list of attributes that better describe the class [59]. The presented approaches aim to improve the model's predictive performance by allowing meaningless relations to remain a part of the model, as long as they positively influence the quality. The resulting models are not acceptable when the task is domain analysis.

In contrast to the presented approaches, we propose a method that constructs multiple models (for example, decision trees) in an algorithmic manner, in order to examine the models' relations for credibility. In this process, a human observes relations in different contexts and, based on common sense, informal knowledge about the domain, the observed relations' frequency, and the stability and quality of the models in which the relations appear, concludes which relations are credible and which are not. Credible relations are then extracted through a specifically designed relation-extraction scheme for overall conclusions. In parallel, *credible models*, composed of credible relations, are also extracted.

Our method, which we have named Human–Machine Data Mining (HMDM), combines DM and human knowledge with the main motivation of transforming the rather ad-hoc process of DM into a systematic procedure. The primary purpose of HMDM is a domain analysis that increases the human understanding of the domain; therefore, HMDM aims to improve the process of finding, not just the models with the best predictive performance, but credible relations and models. The information that the human gathers from the process is just as valuable as the information gained from the results.

The benefits of using HMDM are as follows:

- Humans contribute their knowledge from the real world compensating for incomplete data.
- Computers contribute their immense computing speed and generate new hypotheses under human supervision.
- Humans contribute by directing computer searches into interesting hypothesis spaces.
- Humans perform in-depth analysis on existing data and design superior models in the real world.

The contributions of this paper are: (1) a novel IML method (HMDM) for extracting credible relations and models from data, based on an interactive and iterative process that exploits the advantages of humans and machine algorithms; (2) an extension of the corrected class probability estimate statistical measure, originally conceived for classification rules, that allows it to work on decision trees; (3) interactive explanations of DM results, conceived to facilitate the extraction of credible relations and models; (4) a demonstration of the HMDM method on two real-life domains; (5) an evaluation of the interactive method through a user study.

The remaining part of the paper is structured as follows. The relevant literature is reviewed in Section 2. The interactive method (HMDM) for the extraction of credible relations and models is defined in Section 3. Section 4 demonstrates the method in the two complex macroeconomic domains. The method is evaluated in Section 5. Section 6 concludes the paper and proposes future work.

2. Literature review

In recent years, several methods have been designed to improve the cooperation between humans and DM methods in interactive car systems [48], video retrieval [38], e-mail categorization [17,59], object recognition [18], sensory recommendations [13], and citizen-science projects [31]. The literature indicates that involving a human in DM can significantly improve the obtained results. For example, Kapoor et al. [29] proposed an interactive method for the iterative readjustment of a kernel classifier's decision boundaries, which outperformed automatic DM on three UCI real-life data sets: ionosphere, sonar and heart. Mirchevska [45] showed that increasing human input increases a classifier's predictive performance; the best performance of the study was achieved in the fall-detection domain, for which the amount and quality of human-provided domain knowledge was the highest. Stumpf et al. [59] presented a user co-training approach that combines human- and automatically selected attributes. The resulting approach outperformed the automatic co-training approach.

HMDM incorporates graphs inspired by a research field of explanations dealing with knowledge representations that facilitate the interpretation and evaluation of DM results. The explanations are generally divided into *model-based* [3,5,8,14,24,46,62] and *instance-based* [34,56,60] explanations, depending on whether the goal is to interpret and evaluate the model or to interpret and evaluate the model's prediction.

The majority of explanations start with the assumption that the underlying model does not contain less-credible relations. When this assumption does not hold, the human may draw incorrect conclusions from the presented explanation. For this reason, Stumpf et al. [59] proposed an *interactive explanation* that provides the tools for a human to improve a model, which Kulesza et al. [35] implemented. However, the presented interactive explanations are instance-based and oriented towards improving the model's predictive performance. Consequently, the underlying model can contain meaningless relations that only slightly improve predictive performance. Our task is different. Since our task is domain analysis, we propose using interactive, model-based explanation structures to represent multiple models arranged in a specific manner. This facilitates the extraction of credible relations and models that are both meaningful and of high quality.

Cao et al. [7] grouped the approaches incorporating a human's knowledge into the DM process under the term "domain driven DM". Such approaches typically collect a human's knowledge using some sort of interactive explanation, which enables the human to directly or indirectly modify the model constructed by DM. The modifications are then translated into a computer-understandable form of new training examples, labels for instances, a subset of attributes or constraints imposed on future DM steps. For example, Fails and Olsen [18] and Žnidaršič and Bohanec [74] proposed approaches, in which the human refines the model by providing new examples. MacKay [39], Cohn et al. [11], Tong and Koller [61] and Melville et al. [44] presented an active learning approach, in which the system iteratively asks the human to provide a label for the most informative instance until a satisfactory prediction performance is reached. Ware et al. [68] and Zhao and Yao [72] described approaches for an interactive construction of decision trees that enable the human to test different attributes at each tree-construction step. Stumpf et al. [59] presented a user co-training approach, in which two Naïve Bayesian classifiers—one representing the DM system's view of the data and the other representing a subset of the attributes and their weights as set by the human—are co-trained. Culotta et al. [15], Shilman et al. [58] and Huang and Mitchell [25] looked at the DM process as an optimization process, in which the human's knowledge is incorporated in the form of constraints. Although these approaches propose a variety of modifications, their main goals are the same: to improve prediction by refining a single model. Moreover, they are not intended for domain-analysis tasks.

A typical domain-analysis approach, demonstrated by Nguyen et al. [49] and Osei-Bryson [51], is to: (a) construct multiple models by varying the DM method's parameters, (b) preselect models based on formal criteria, and (c) enable the human to select one or several of the models based on informal knowledge. Since the models are automatically constructed, they can contain less-credible relations, as presented in the introductory example. To eliminate less-credible relations from the analysis, we propose a new method, which upgrades the presented approach with attribute- and constraint-based modifications typical of prediction tasks. In this manner, the relations in the models are examined for credibility, and only the relations and models that prove to be credible are extracted.

Interactive approaches similar to ours are presented by Maeno and Ohsawa [40] and Kuntz et al. [36], who consider the human to be a heuristic that guides the DM system towards association rules that are supported by the data and are, at the same time, interesting for the human. In contrast, our method deals with classification and regression rules. Moreover, our types of interaction and explanation differ from those of the related approaches.

Martens and Baesens [42] and Martens et al. [43] proposed a measure of justifiability that indicates the degree to which an automatically constructed DM model is in line with existing domain knowledge. Based on a decision table constructed from the model, the human expresses views of a domain by stating levels of attribute importance as weights. Inconsistencies between the model and the human's view are penalized more for attributes with higher weights, thus reducing the model's level of justifiability.

Feelders [19] selected monotonic decision trees as being in line with human expectations in economic domains. Multiple trees were constructed from different random training and test data partitions in order to construct the monotonic decision trees. The interest was primarily on estimating the percentage of monotonic trees constructed with an automatic decision tree learner. In contrast, our approach constructs multiple models in a methodological manner to observe the interesting relations in different contexts and to conclude which relations are credible.

Some procedures of HMDM are based on attribute selection mechanisms. One method that has a similar purpose is Wrapper Attribute Selection (WAS) [33]. WAS aims to find the optimal subset of attributes for a specific DM method and a specific data set, such that the prediction model constructed from these attributes is of the highest possible quality. To achieve this goal, WAS conducts a search using the DM method as part of an evaluation function. In Section 5, we compare WAS with the attribute selection mechanisms of HMDM.

3. Human–machine data mining

According to constructivist learning theory, the best learning framework for a human is experimental task-based learning, which enables the human to actively integrate the acquired knowledge into his/her existing mental model of the domain or to use the acquired knowledge to challenge and modify his/her existing model. Whereas a single model constructed with classic DM is comparable to a teacher delivering a didactic lecture covering the subject matter, the IDM method that we propose is comparable to a facilitator helping a learner to understand the content [37]. A corresponding scheme is presented in Fig. 3.

The basic idea of our approach is to construct multiple interesting models and relations from the data that can be accepted by the human as meaningful and of high quality and, therefore, credible. Without the method for systematically constructing, examining and extracting relations and models, a human would need to examine the search space of models in a rather ad-hoc manner to find those that are credible. Since there are 2^{2^n} possible models for n binary attributes and a single binary class, for n higher than four it would be very difficult for a human to analyze all hypothesized models. Therefore, some method is needed to select a reasonable number of interesting hypotheses from all of the possible options. The method we propose combines human understanding and raw computer power to enable a smart examination of the credible relations and models within the huge search space of models. Through our approach, when DM methods perform a search, humans examine and evaluate the results, make conclusions and redo the search in the way that seems most promising, based on the previous attempts. In this way, the humans guide the DM to search the subspaces with the highest probability of credible relations and models. Finally the humans, in their minds, construct the overall model of the domain using the most interesting conclusions.

Our approach is based on the assumption that HMDM will help discover the credible relations and models from the enormous number of possible models, despite humans' subjectivity in terms of preference. Indeed, this is our experience in recent years in most real-life domains. The novelty is that our approach transforms the rather ad-hoc process of DM into a systematic heuristic procedure.

We use two basic heuristics: First, we examine the whole set of various parameters (typically, the DM method's parameters and attribute subsets) to obtain information about where the credible relations and models might be, and second, as soon as a candidate model occurs, we apply several heuristics to cross-check the credibility of the observed and similar relations and models.

The cross-checking heuristics are based on two ideas. First, if a specific relation in a model is confirmed as credible, the relation is added to the list of candidates for credible relations. When all relations from the candidate model are confirmed as credible, the model is added to the list of candidates for credible models. Second, the credibility of a relation can be confirmed by deleting and attaching attributes. In other words, when the attributes that form a relation are deleted from the data set and the resulting model (reconstructed from the reduced data) is of lower quality, the relation gains in credibility, and vice versa. Similarly, when the attributes that form the relation are attached to a set of attributes S and the quality of the model constructed from the resulting set exceeds the quality of the model constructed from S, the relation gains in credibility. Relations and models that are both supported by the evidence and accepted by the human are stored in the candidate lists, which are refined in new search cycles that are repeated until no new, interesting relations are found and no new evidence is observed that will confirm or disconfirm the credibility of the candidate relations and models.

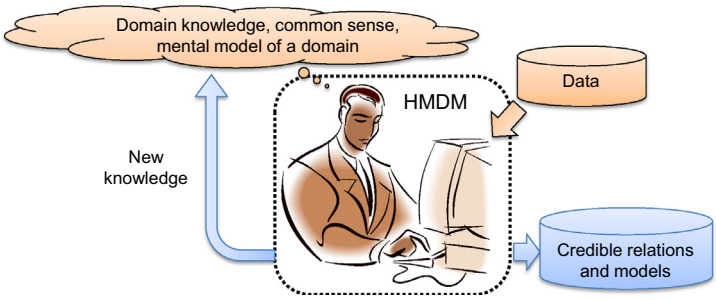

Fig. 3. Interactive data mining schema: A human and a computer cooperate to improve the human's knowledge.

3.1. Illustrative examples

The idea of the proposed approach is illustrated with two examples. The first example shows how to determine a relation's credibility by attaching attributes, and the second example shows how to determine a relation's credibility by deleting attributes.

3.1.1. Attach attributes

Let there be a data set D, presented in the table in Fig. 4, which contains three binary attributes, A_1, A_2, A_3, and a binary class C. Then, J48 decision-tree models are constructed from D. First, several trees are constructed from all of the attributes with different parameters of the J48 algorithm. From the constructed trees, the tree in the bottom-left corner in Fig. 4 is selected as a candidate tree. The credibility of the relations contained within the candidate tree is examined through the construction of additional trees. The constructed trees are arranged in an interactive explanation structure called an *attached attributes graph*, an example of which is presented in the central part in Fig. 4. The root of the graph represents an empty attribute set (NO ATTRIBUTES). The first node in the graph (A_1 | 71.43) represents the tree constructed by attaching the attribute A_1 to the empty set. This tree constructed only from the attribute A_1 is presented in the top-middle part of Fig. 4. The quality of this tree, expressed as the ACC, is 71.43%. The second node of the attached attributes graph represents the tree constructed by attaching two attributes (A_1 and A_2) to the empty set. This is indicated by the hierarchical structure that connects the two attributes in the graph. This time, the ACC of the corresponding tree is 100%. The third node in the graph represents the tree constructed by attaching the attribute A_2 to the empty set. Additional nodes that represent the trees constructed from other combinations of attributes can be added to the graph; however, they are omitted here in order to simplify the example.

The presented graph contains one good candidate for a credible relation, which is marked in bold. Its credibility is established by comparing the ACC of the trees constructed from A_1, A_2 and the combination of the two. Considering that the ACC of the combination (100%) is considerably higher than the ACC of the other two trees (71.43% and 85.71%), the next step is to examine the tree constructed from the combination of attributes for meaning. If the tree is meaningful to the human performing the examination, a ***combination relation*** is established, which is denoted as A_1 & A_2. Although the relation is supported by the evidence, it has not been included in the candidate tree. The relation is added to the list of candidates for credible relations, and this list is further reexamined for credibility through additional analysis steps, as explained in Section 3.2. The credible model constructed from the two attributes is stored as well.

3.1.2. Delete attributes

In this example, J48 trees are constructed from a data set D', presented in the table in Fig. 5, which contains three binary attributes, A_1, A_2, A_3, and a binary class C. This time, the trees are arranged in an interactive explanation structure named the *deleted attributes graph*, which is presented in the central part in Fig. 5. The root of the graph represents the candidate tree constructed from all three attributes (ALL ATTRIBUTES | 100). The candidate tree presented in the top-right corner of Fig. 5 indicates that a single important relation is the relation between the attribute A_3 and the class. For the relation to be credible, a new tree constructed from the data set obtained by deleting A_3 from D' should have a lower ACC. However, when A_3 is deleted, the ACC remains the same (100%). In the new tree, it can be seen that another attribute, A_1, took the role of A_3 in the tree. When two attributes are semantically similar—for example, when both attributes represent the level of investment in R&D, but one is expressed as "per capita" and the other as a "percentage of GDP"—it indicates a ***redundancy relation***. To confirm the redundancy relation, the next step is to delete both attributes from D' and to construct a tree from the reduced data set. Since the ACC of the new tree falls (to 71.43%), the redundancy relation $A_1 || A_3$ is established and added to the list of candidates for credible relations.

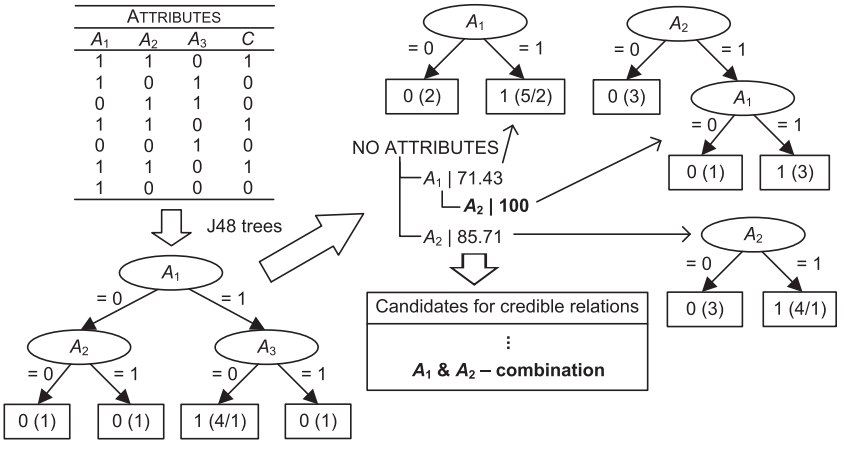

Fig. 4. Example of a combination relation established for the data set D by constructing J48 decision trees.

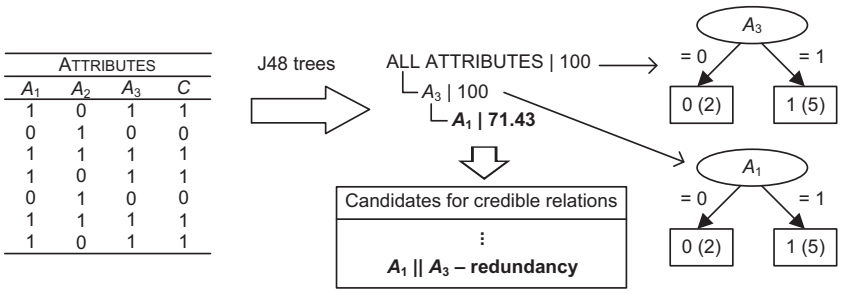

Fig. 5. Example of a redundancy relation established from the data set D' by constructing J48 decision trees.

These illustrative examples presented the procedures for establishing combination and redundancy relations by attaching and deleting attributes. Combination and redundancy are two *types* of credible relations that can be extracted through the relation-extraction scheme we propose. An additional mechanism for determining relations' credibility attributes the *levels of credibility* to relations and represents the second dimension of the relation-extraction scheme, which is appropriately named the ***type-credibility scheme***. The additional mechanism is described next.

3.2. Credibility

The second dimension of the type-credibility scheme attributes one of the three levels of credibility to a relation. The level is determined by observing the relation's *frequency* and *stability* through multiple models within the attached and deleted attributes graphs. The ***first-level*** relation frequently appears in models and is stable. Its stability is determined by observing changes in its quality, caused by the presence or absence of a relation in the data. In other words, the deletion of attributes from a stable relation generally causes a fall in the quality of the models constructed from the reduced data, and vice versa. The ***second-level*** relation is frequent, but less stable. For example, the deletion of attributes from a less-stable relation causes a fall in quality in one case and an increase in quality in another case. Second-level relations frequently describe subgroups within the data. The ***third level*** contains all the other relations, which are of low quality and/or for which the presence or absence of relations in the data does not affect the quality. These are, consequently, less interesting for the analyst. Because the nodes in the graphs indicate the levels of credibility, they are referred to as *credibility indicators*.

Based on the proposed relation-extraction scheme, we derive the following definitions:

Definition 3.2.1. *Meaning* is a subjective criterion attributed by the human based on common sense, informal knowledge about the domain, observed frequency and the stability of a relation. A relation can be either *meaningful* or *meaningless*. The boundary between the two states is human-defined, but based on observed evidence.

Definition 3.2.2. *Quality* is an objective criterion that indicates the support of the selected quality measures. Quality is expressed in both numeric and discrete form. In the latter form, the quality can be either high or low. A high-quality model has a higher quality than that of the baseline, while a low-quality model has a quality that is equal to or lower than that of the baseline. The baseline is domain-dependent. A high-quality relation appears in the high-quality model, and a low-quality relation appears in the low-quality model.

Definition 3.2.3. *Credibility* is a property of trustworthiness. A credible relation is both meaningful and of high quality. A less-credible relation is either meaningless, of low quality or both. A less-credible relation is less often confirmed in the search process. Relations that are confirmed more often and more strongly are more credible. A credible model is composed of credible relations.

The motivation for defining the types and levels of credibility is similar to the motivation behind the multi-view approach to intelligent data analysis [9], which is to enable different users to view data from different perspectives in order to find the relations that best describe the domain in terms of both quality and meaning.

3.3. The algorithm

The credible relations and models are extracted with the help of the type-credibility scheme using the following procedure:

(1) The human begins the analysis by supplying the data of interest,
(2) Within the loop, the human:
  (a) initializes the search by finding an initial model,
  (b) interactively searches the domain around the initial model, guiding the search with the help of:

- DELETE_ATTRIBUTES and ATTACH_ATTRIBUTES procedures and the expand credibility indicator tool,
- several quality measures (Section 3.4) and the heuristic function in Eq. (1).

We call this method Human–Machine Data Mining (HMDM). It is based on a typical DM search, consisting of choosing an interesting subset of attributes and then performing DM with various parameters. However, rather than ending with the best model, as is common in a DM search, HMDM relies on human decisions to choose one or a couple interesting candidate models, then re-analyzes them by changing the parameters and attributes in the prescribed manner and repeats the loop until the relations within the model(s) are confirmed or rejected. Too many variations of search options cause a combinatorial explosion; however, the search is guided by a human's goal: to verify the already-found interesting relations. As seen in the examples in Section 4, several variations can quickly be discarded as unpromising, and the interesting ones can receive human attention, demanding only reasonable time consumption.

The steps of the HMDM algorithm are as follows, with human parts denoted with (**H**) and computer parts with (**C**):

Step 1 – Select a data set: **H:** The human selects a data set that describes a domain of interest.

Step 2 – Modify the attribute set: **H:** The attribute set modification step is omitted when the data is analyzed for the first time. During the first cycle of analysis, the human may hypothesize that certain attributes expressed in a different form may contribute to domain understanding (for example, a new attribute indicating the sector that invests the most in R&D, rather than six attributes indicating the percentage shares of the total investment made by the six sectors). In the second cycle, the human constructs new attributes to test the hypotheses by applying sum, min, max or ratio functions on two or more numerical attributes.

In contrast, when the human establishes that certain attributes are not interesting for further analysis, he/she eliminates those attributes by applying different forms of granulation [2].

C: When the human decides that a too-large number of attributes disrupts the analysis, he/she executes an arbitrary automatic attribute subset selection method.

Step 3 – Select DM method: **H:** The human selects one or several of the DM methods designed for supervised learning problems, which are capable of producing models in a human-understandable form (the HMDM method was tested and verified on decision and regression trees; the examples are presented in Section 4).

H, C: For any chosen DM method, the human–computer session is performed until the human explores all the interesting models, meaning that no new high-quality models containing meaningful relations are constructed.

Step 4 – Select parameters and their ranges, define constraints: **H:** By defining the parameters and constraints, the human defines his/her preferences regarding the hypothesis subspace searched by the DM method. For example, the subspace may consist of all the decision trees containing two or more relations.

Step 5 – Initial DM: This step represents an exploratory data analysis phase [41].

C: The INITIAL_DM procedure constructs all the models possible with the selected DM method and the defined parameters that satisfy the constraints. When multiple models are constructed, they are sorted using the predefined criteria. To reduce information overload, additional instances of the same model are removed. If no models are constructed, the human is redirected to Step 3 to select another DM method, under the assumption that another knowledge representation would be better suited to the domain. If all DM methods fail to construct models, the analysis is terminated. This step resembles the construction part of the random forest algorithm [6], with the difference being that the human defines the way in which the new models are constructed.

H: Once they have been constructed, the human examines the models in decreasing order of sorting criteria and selects one or several interesting models for further analysis. Here, the human may even select a low-quality model if he/she wants to obtain additional evidence regarding whether the relations in the model are credible.

Step 6 – Interactive domain search: Each model marked as interesting is a starting point in the search for credible relations and models. The searching steps of deleting and attaching attributes are separately executed for each model, which takes on the role of the initial model during these steps.

C: In this process, the first step is the execution of the DELETE_ATTRIBUTES procedure. This procedure constructs a deleted attributes graph by first deleting attributes from the initial model and then deleting attributes from the models constructed from the reduced attribute sets. An attribute is added to the graph when: (a) the quality of a model M' constructed after eliminating the attribute (using INITIAL_DM and selecting the top-ranked model) is lower than the quality of the superordinate model M and (b) the structure of M' is different from the structure of M. In other words, the procedure terminates the process of adding new attributes to the graph's depth when the first attribute that either improves the model's quality or reduces its quality but does not change its structure is encountered. The procedure resembles WAS with a backward elimination approach [33]. The main difference is that our procedure returns a human-readable graph that contains multiple models and relations constructed from different attribute subsets. The human uses the graph to examine the relations and models for credibility, considering not only their quality, but also their meaning. In contrast, WAS outputs the highest-quality model found, while hiding the models and relations constructed in the process from the human.

H: If the human hypothesizes additional interesting relations that are not included in the graph, the human refines the graph by applying the expand credibility indicator tool. The human examines the constructed graph, extracts credible relations and models with the help of the type-credibility scheme and adds them to the list of candidates for credible relations and models.

C: The candidate relations are further examined by attaching the attributes from the relations. This is achieved by executing the ATTACH_ATTRIBUTES procedure, which systematically constructs credibility indicators in reverse order from that of DELETE_ATTRIBUTES, starting with an empty set and then gradually attaching attributes from the subset. The main difference between the two procedures is that the ATTACH_ATTRIBUTES procedure adds a new attribute to the graph when the quality increases after attaching the attribute, and the attached attribute is present within the model's structure. The procedure resembles WAS's forward selection approach [33], with the main difference being in the termination condition. WAS terminates the search when a new attribute does not increase the quality of the constructed model. In contrast, our procedure will also terminate the search when the quality improves, but the model's structure—that is its meaning—remains the same. This occurs when the attached attribute appears in models constructed from cross-validation folds, but not in the model outputted by the DM method. Humans find such cases confusing and lacking in added value for further analysis.

H: When the human hypothesizes additional interesting relations, he/she also reexamines them, either by selecting a specific subset of attributes to serve as an input to the automatic procedure or by refining the graph with the help of the expand credibility indicator tool.

H: To confirm or reject the established type of relation, the human applies the heuristic function in Eq. (1), called "interaction", proposed in Jakulin [26].

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

S represents a set of attributes from the relation to which the class is added. $P(v)$ represents a (joint) probability distribution for values of attribute(s) within the subset $T \subseteq S$. A positive interaction indicates a synergy between the attributes, that is, a combination. A negative interaction indicates an overlap between the attributes, that is, redundancy.

Step 7 – Store credible relations and models, integrate conclusions: **C:** The relations and models confirmed as credible by the human are stored.

H: The human integrates the conclusions based on the credible relations and models with the conclusions made from previous analyses of the same data.

We should note that an automatic attribute construction method is not used in Step 2, since (a) the human is focused on those combinations of attributes that he/she finds interesting from the results of previous analyses; (b) this is a hypothesis testing step, not a data exploration step [41], and it is not relevant to bring additional combinations to the human's attention that do not bear meaning for the human (for example, the total number of female researchers and the percentage of female researchers will not offer any additional insight into an analysis regarding how various segments of the R&D sector influence a country's economic welfare); and (c) the literature suggests the construction of domain-specific attributes in the presence of domain knowledge [22].

3.4. Quality measures

Comparisons between the models within the deleted and attached attributes graphs are based on several quality measures. First, for each group of the classification and regression models, we implement two standard measures of quality that are applicable to a wide spectrum of models. Second, we propose a measure of the corrected class probability estimate (CCPE). The measure was originally conceived to operate with classification rules [47], but we adjusted it to operate on decision trees. Note that the CCPE is applicable to transparent-box models other than decision trees. However, in this paper, we present only the adjustment for decision trees. Finally, we propose a q_Δ measure that connects the presented measures into a single, quality-based comparison criterion.

3.4.1. Standard measures

The quality of the classification models is estimated with two standard measures: ACC (see Section 1) and Kappa. Cohen's Kappa indicates whether the agreement between the classifier's predictions and the actual class values exceeds the chance level [10]. In theory, the values of Kappa are distributed within the $[-1, 1]$ interval, where values less than or equal to 0 indicate a random classifier and a value of 1 indicates the best classifier. In practice, we did not consider the random classifiers with negative Kappa.

The quality of the regression models 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" [69]. A CC of 0 denotes the worst, while a CC of 1 denotes the best model. 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 [69] from 100, such that higher values denote a better model.

3.4.2. Adjustment of the CCPE

The CCPE reflects the decision tree's quality in comparison to the qualities of all possible decision trees constructed from the data. The values of the CCPE are distributed within the [0, 1] interval, with 0 indicating the worst and 1 indicting the best decision tree.

The CCPE corrects the standard class probability estimate (CPE), which is prone to assigning optimistically high values to random patterns in the data. This optimism is reduced by subtracting the proportion of rules with qualities higher than that of 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 or rules. For each relation, the probability of the majority class in the leaf (that is, the CPE) is computed first. The value of the CPE is then reduced by the proportion of the relations whose qualities are higher than that of the examined relation. For this purpose, we compute the Fisher–Tippet extreme value distribution (EVD), which represents the qualities of all possible relations that can emerge in decision trees constructed from the data. The distribution parameters are computed from a random sample of decision trees constructed from the data, which means that they are estimates and cannot guarantee that all the possible trees have been considered. Finally, the CCPE for each relation is weighted by the proportion of examples covered by the relation and then summed to obtain the CCPE for the tree. The described procedure is composed of two parts: one for computing the EVD and another for computing the tree's CCPE.

The procedure for computing the μ parameter of the EVD is presented in Table 1. The EVD has two parameters: μ, representing location, and β, representing scale. The statistic sampled to compute the EVD is a log-likelihood ratio statistic (LRS), computed using an equation from the fourth line in Table 2. For the LRS, β is always equal 2. Therefore, the procedure computes only the μ parameter, which is dependent upon the maximum depth of the tree. For each depth, μ is computed by constructing a predefined number of trees (in our case, 1000) from the given data, 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 procedure are constructed using the modified J48 algorithm from Weka, which we modified to: (a) use the LRS instead of the entropy and (b) complete the construction of the tree at the predefined maximum depth.

The procedure for computing the tree's CCPE is presented in Table 2, in which s denotes the number of the majority class c 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 the examples that did not reach the leaf.

3.4.3. q_Δ measure

Within the deleted and attached attributes graphs, the effect of deleting and attaching attributes is assessed through the summary measure q_Δ. Let ACC_Δ be the difference in the ACC between a classification model constructed by deleting an attribute 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:

$$q_\Delta = (ACC_\Delta/100 + CCPE_\Delta + Kappa_\Delta). \tag{2}$$

Table 1
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 → Dp
      Learn a decision tree on Dp with LRS as an 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 μ(d) > μ(d − 1)
  Return the list of μ parameters for different depths
```

Table 2
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 = 2[ s log(s/e_s) + (n − s) log((n−s)/e_{n−s}) + s^c log(s^c/e_{s^c}) + (n^c − s^c) log((n^c−s^c)/e_{n^c−s^c}) ]
    Compute area P under the EVD(μ(relation depth), β = 2) with the LRS as a lower bound
    Compute expected value of LRS (LRS~) by finding the lower bound for the area under χ²(1) equal P
    Compute the expected value of s(s~) from the LRS~ using a root-finding algorithm
    CCPE = CCPE + (s~/n) × (n/(n+n^c))
  END FOR
  Return CCPE
```

Expected values: $e_s = n\frac{s+s^c}{n+n^c}$; $e_{n-s} = n \times \left(1 - \frac{s+s^c}{n+n^c}\right)$; $e_{s^c} = n^c\frac{s+s^c}{n+n^c}$; $e_{n^c-s^c} = n^c \times \left(1 - \frac{s+s^c}{n+n^c}\right)$.

By dividing with 100, we reduce the ACC to the same scale as that of the CCPE and the Kappa, so that all measures are equally weighted. The interpretation of the values for q_Δ differs based on whether the attributes are being deleted or attached and is prescribed by the type-credibility scheme. Similarly, for regression models, q_Δ represents a linear combination of CC and RAA, where RAA is divided by 100.

In summary, the HMDM algorithm contains the following components:

(1) the collection of the DM and ML algorithms, such as Weka or Orange [16], used for constructing models from data;
(2) procedures: Initial_DM, Delete_attributes, Attach_attributes, Compute_EVD and Compute_CCPE;
(3) tools: the expand credibility indicator;
(4) standard routines, such as attribute selection;
(5) quality measures and the heuristic function.

The HMDM algorithm is available for academic purposes as a research software tool at http://dis.ijs.si/Vedrana/HMDM.htm, where you can also find the flowchart of the HMDM algorithm.

4. Real-life examples

This section presents two applications of the HMDM method on the real-life domains of higher education and R&D. First, the domains and their corresponding data sets are introduced. Second, the experimental setup is described. Third, the step-by-step example of applying HMDM to the higher education domain is presented, followed by the conclusions drawn from the analysis. Finally, we present the conclusions from the analysis of the R&D domain.

4.1. Domains

4.1.1. Higher education domain

The aim of this analysis was to understand which segments of the higher education sector have an impact on the economic welfare of a country. To accomplish this task, we collected data for the year 2001, composed of 60 attributes representing the higher education sector and 167 examples representing countries from two statistical databases provided by the UNESCO Institute for Statistics (http://www.uis.unesco.org) and the USAID Global Education Database (http://ged.eads.usaidallnet.gov). From these data, two data sets were created: HI-EDU$_{class}$ and HI-EDU$_{reg}$. The former was intended for the construction of the classification and the latter for the construction of the regression models. In both data sets, the economic welfare was represented with the class *GNI per capita*, where GNI stands for the Gross National Income and denotes the total value of goods and services produced within a country [4]. The class for HI-EDU$_{class}$ was collected from The World Bank database (http://www.worldbank.org) in a discrete form that represents the official classification of the countries into income levels: low—745 US$ per capita or less (50 countries); middle—746–9205 US$ per capita (79 countries); and high—9206 US$ per capita or more (38 countries). The class for HI-EDU$_{reg}$ was obtained by transforming the discrete class into a numerical representation by encoding low as 1, middle as 2 and high as 3. Within the preliminary research [64], we experimented with the class stated in US$; however, the resulting regression trees were of poor quality (CC of 0.52).

In the advanced stages of the analysis, modifications steps (corresponding to Step 2 of the HMDM algorithm) were applied to the two data sets, resulting in HI-EDU$_{class\text{-}mod}$ and HI-EDU$_{reg\text{-}mod}$. The former was obtained by constructing nine attributes based on the observations and deleting twenty attributes differentiating between the female and male students (since the better status of females in higher education is a consequence of better economic welfare). The latter data set was obtained by deleting twenty female-male status attributes.

For the purpose of testing HMDM, we also collected data for 2010. The same set of 60 attributes was collected, and the same modification steps were applied. After examples with significant proportions of missing values were eliminated, the final 2010 data sets comprised 125 examples.

4.1.2. R&D domain

The aim of this analysis is to understand which segments of the R&D sector have an impact on the economic welfare of a country. For this purpose, we collected data for the year 2001, composed of 48 attributes representing the R&D sector and 167 examples representing countries from two statistical databases provided by the UNESCO Institute for Statistics (http://www.uis.unesco.org) and WIPO (http://www.wipo.int). Two data sets were created: first, for the discrete GNI per capita class (R&D$_{class}$) and, second, for the numeric GNI per capita class (R&D$_{reg}$), this time expressed in US$ (source: The World Bank). Modifications of the two data sets (R&D$_{class\text{-}mod}$ and R&D$_{reg\text{-}mod}$) were obtained by constructing 19 attributes.

Data sets for 2010 were also created. These had the same structure as the data sets for 2001, but contained 78 examples.

The presented data sets, a description of the attributes therein, and a description of modifications are available at http://dis.ijs.si/Vedrana/economic-analysis.htm.

4.2. Experimental setup

The HMDM method was applied in combination with both the classification and regression DM methods. The classification models were constructed using the J48 and regression models using the M5P algorithm [55] from Weka. For the J48 algorithm, two parameters that control the model's complexity were selected: MNIL, with values ranging from 2 to 15, and reduced-error pruning (REP), with on/off values. For the M5P, only the MNIL parameter was selected, with values ranging from the default 4 to 15. For parameter values other than those stated, the default values set in Weka were used.

Separate analyses were conducted for 2001 and 2010 by two independent persons. Both humans executing HMDM for a particular year possessed general knowledge of a knowledge representation constructed by a selected DM method (a decision or regression tree) needed to interpret the model and to extract relations. Each human first extracted candidate relations by observing the quality, frequency and stability of the relations. Then, each human refined the set of candidates using domain-related literature.

The quality of the constructed models was computed using a 10-fold cross-validation [32] with a random seed equal to 1.

For the one-decade-apart comparisons (2001 to 2011), there were insufficient data available on the Internet to enable fair comparisons.

4.3. Analyses with the higher education data

4.3.1. An example of constructing decision trees from the higher education data

The INITIAL_DM procedure was applied to the HI-EDU$_{class}$ data set with the parameters defined in Section 4.2. From the 14 decision trees returned by the procedure, we selected the tree constructed with the parameter MNIL 7. The tree was of the highest quality (ACC 71.86%; CCPE 0.5879; Kappa 0.5497) but, at the same time, contained relations meaningless in the context of the analysis. For example, consider a relation indicating the importance of the percentage of students graduating in agriculture: First, based on our common sense, we doubted that the relation was important for a country's welfare (since we would expect that the percentage of graduates in science would have a higher impact); second, the relation did not make a clear distinction between countries of different levels of welfare.

4.3.1.1. Delete attributes. In order to examine the credibility of the relations in the initial tree, we began the analysis by constructing a deleted attributes graph: first, by applying the automatic DELETE_ATTRIBUTES procedure, and second, by refining the graph with the credibility indicator tool to test additional hypotheses. In the interactive approach, the human directed the search of DM methods towards the parts of search space not considered by the automatic DELETE_ATTRIBUTES procedure, but still containing evidence that could support or reject interesting relations and models. For example, the nodes below the fourth node in Fig. 6a (the root is not counted as a node) are added interactively to verify discovered relations. Note that the complete graph is larger, but that we have presented only one part with interesting findings.

The graph is presented in Fig. 6a. The root node represents the initial tree, which is constructed from all attributes. The numbers divided by the vertical bar represent the ACC, CCPE and Kappa of the initial tree. The first node represents a tree constructed from the reduced data set, which is obtained by deleting the attribute "Gross outbound enrolment ratio" (GOER) extracted from the initial model. The numbers in brackets that follow the quality measures serve to track the changes in quality between the current tree and the initial tree. The last number represents q_Δ, which indicates the total change in quality between the two trees. By observing the node, we can extract the first candidate for a credible relation, which indicates that the mobility of students (GOER) is important to a country's welfare. The evidence supporting the relation is the following: When the GOER attribute is deleted, all three quality measures fall (as indicated by the negative values in brackets), resulting in a q_Δ of -0.0741. The node representing the relation is marked by HMDM (based on an objective quality criterion) in bold to indicate the candidate for the credible relation. The second node represents a tree constructed by deleting both the GOER and the "Public expenditure on education as % of GNI" (PE-GNI) attribute. Here, a hierarchical structure represents a deletion of multiple attributes, and the second number in brackets represents the difference between the current tree and the superordinate tree. For the second node, we can observe a negligible fall in quality (q_Δ of -0.0779) in comparison to that of the superordinate tree (q_Δ of -0.0741), which may indicate a redundancy relation. To confirm the redundancy, we visually observe the tree represented by the node to discover whether it contains an attribute semantically similar to PE-GNI. We establish that the attribute "Current expenditure on education as % of GNI" (CE-GNI), which, indeed, represents the same semantic category takes the place of PE-GNI in the tree. Both attributes represent the level of investment into all levels of education. While public expenditure is composed of both current and capital expenditures, current expenditures (CE-GNI) form the majority of public expenditures. Considering the lack of ontology describing such semantic relations between attributes, this domain knowledge was incorporated into the DM process by the human. The deletion of the CE-GNI attribute (third node) confirmed the redundancy relation PE-GNI $\parallel$ CE-GNI with a fall in all quality measures. In addition, the relation was confirmed by the negative interaction of -0.049. This relation represents the second candidate for credible relations. Finally, the three nodes—fourth, fifth and sixth—indicate the third candidate. The three attributes beginning with "Gross enrolment ratio. ISCED 5 and 6" (GER) all represent the level of participation in higher education, expressed as total participation, the participation of females and the participation of males, respectively. The three attributes substitute for one another, and when they are deleted together, the fall in quality is considerable (q_Δ of -0.1126). The semantic relation

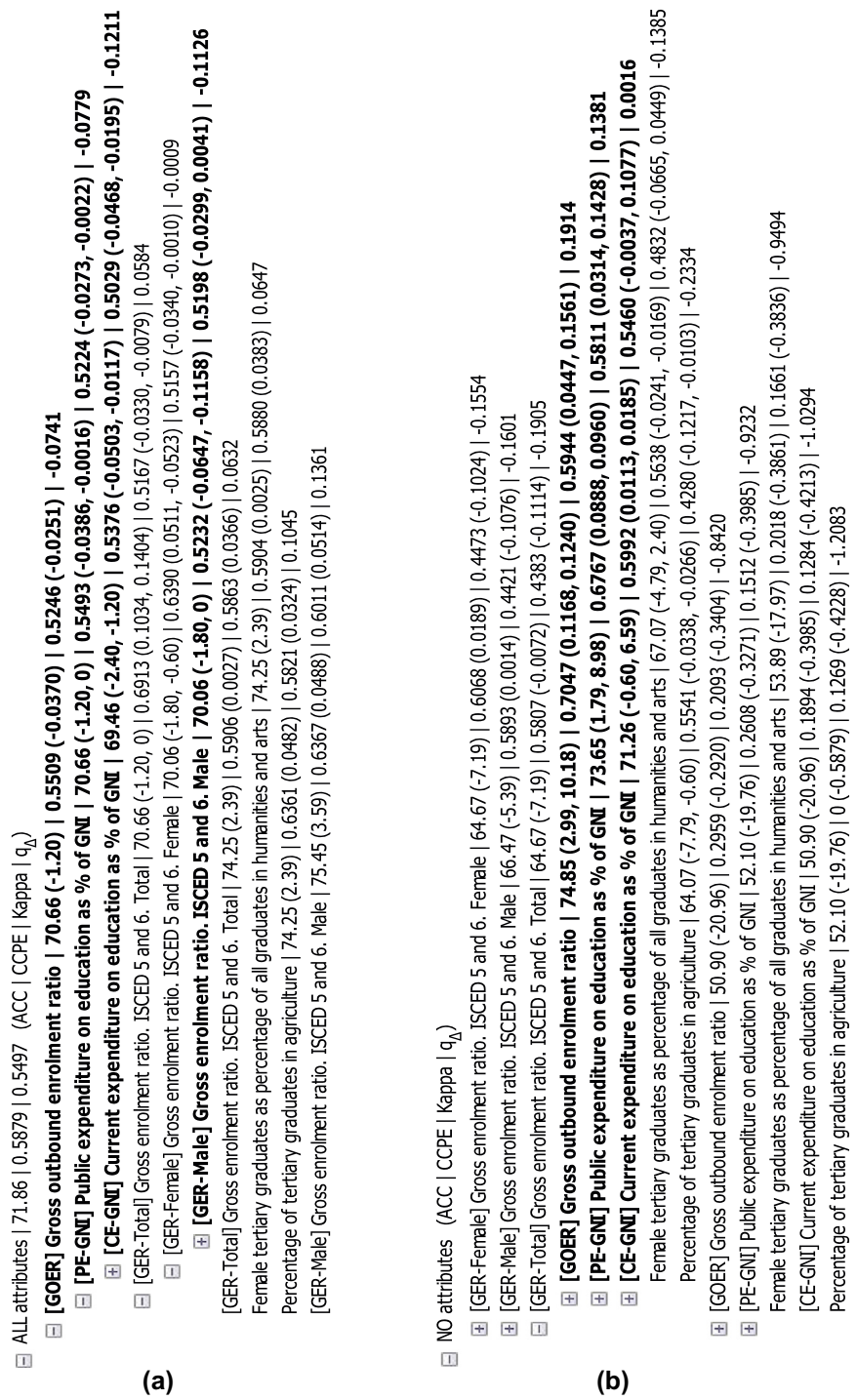

Fig. 6. Graphs constructed from the higher education data: (a) deleted attributes graph and (b) attached attributes graph.

was again established by the human. Therefore, we can establish the redundancy relation GER–Total || GER–Male || GER–Female.

4.3.1.2. Attach attributes. We further reexamined the relations from the list of candidates for credible relations by constructing an attached attributes graph. The graph is constructed with the automatic ATTACH_ATTRIBUTES procedure and refined by the interactive expand credibility indicator tool.

The graph in presented in Fig. 6b. The root of the graph represents an empty attribute set. The nodes at the first level of the hierarchy represent the trees constructed only from the attribute in the node. The nodes at the deeper levels represent trees constructed from the combination of attributes. As in the deleted attributes graph, each node contains the three measures of quality and the differences in qualities. In contrast to the deleted attributes graph, in this graph, the positive differences are preferred.

In the presented graph, only three nodes are interesting for further analysis based on the quality criterion; these are marked in bold by the HMDM. The first indicates the combination of GER–Total and GOER (fourth node): When the tree is constructed only from the GER–Total (third node), the q_Δ is negative (-0.1905); similarly, when the tree is constructed only from the GOER (ninth node), the q_Δ is again negative (-0.842); however, when the tree is constructed from the combination of the two attributes, the q_Δ is positive (0.1914). To confirm the relation, we examined the tree represented by the fourth node (in Fig. 7a) for meaning. The tree contains meaningful relations, which indicates that, for better welfare it is important to stimulate participation in higher education and to improve student exchange programs, especially for those students that leave the country to study abroad. The relations are common sense and are partly supported by the literature (see Section 4.3.2). Based on the presented evidence, the combination relation GER–Total & GOER is established. Based on additional analyses, we determined that the trees of similar structure, semantics and quality are obtained by substituting the GER–Total with any of the GER–Female or GER–Male attributes. Therefore, the relation GER–Total & GOER is translated into the (GER–Total || GER–Male || GER–Female) & GOER relation. Since the relation with its subparts is frequent and stable through both the deleted and the attached attributes graphs, it is attributed with the first level of credibility. In addition, the positive interaction supports the relations of GER–Total & GOER (0.0193), GER–Male & GOER (0.0095) and GER–Female & GOER (0.0037). The second and third interesting nodes support the relation PE-GNI || CE-GNI. Both nodes (fifth and sixth in the attached attributes graph) represent trees of higher quality than that of the initial tree, which is indicated by a positive q_Δ of 0.1381 and 0.0016. By observing the trees, we can see that the two attributes substitute for each other. The tree constructed from the GER–Total and PE-GNI (fifth node) is presented in Fig. 7b. It indicates that for "low" countries to improve their welfare, it is important to increase investment in all levels of education. The other tree, constructed from the GER–Total and CE-GNI attributes (sixth node), differs from the first tree only in the left subtree, where the PE-GNI subtree in Fig. 7b is replaced by the CE-GNI subtree, which divides "low" and "middle" countries in the same manner. This relation is common sense and is supported by the literature (see Section 4.3.2). The relation is attributed with the second level of credibility, since it describes a subgroup within the data.

4.3.1.3. Integrate conclusions. Several hundred additionally constructed trees confirmed the first impression, which was that the most important measure to improve welfare is to stimulate participation in higher education and to improve student exchange programs, especially for those students that leave the country to study abroad: (GER–Total || GER–Male || GER–Female) & GOER. In addition, developing countries should increase their levels of investment in all levels of education: PE-GNI || CE-GNI.

4.3.2. Conclusions from the analyses of the higher education data

Additional analyses were performed by constructing decision trees from the HI-EDU$_{\text{class-mod}}$ data set and regression trees from HI-EDU$_{\text{reg-mod}}$. As a result, we extracted the following credible relations:

- First level of credibility:
  - (GOER || "Outbound mobility ratio (%)") & (GER–Total || GER–Male || GER–Female || "Tertiary students per 100,000 inhabitants"). The most important measure to improve the welfare for all countries is to stimulate participation in higher education and to improve student exchange programs, especially for those students that study abroad.
- Second level of credibility: When the measures belonging to the first level have been taken, the additional measures are:
  - PE-GNI || CE-GNI || "Public expenditure on education as % of GDP". For "low" countries to improve their welfare, it is important to increase the level of investment in all levels of education.

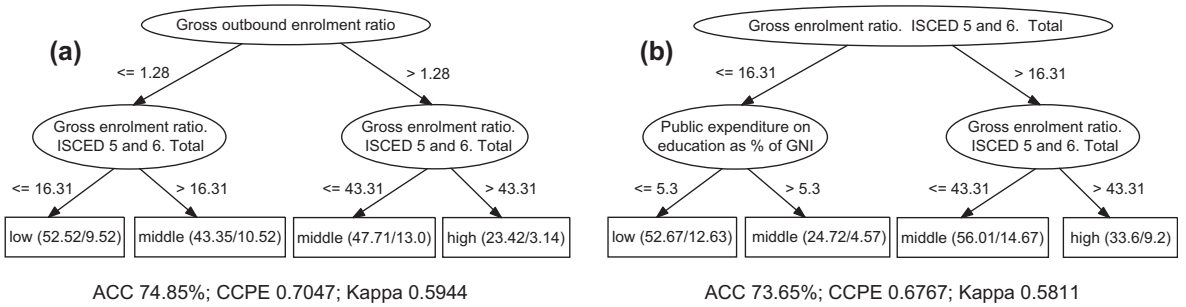

Fig. 7. Two credible trees constructed from: (a) the GER-total and GOER, MNIL 11 and (b) the GER-total and PE-GNI, MNIL 11.

- ''Percentage of tertiary graduates in science''. For "middle" countries to improve their welfare, it is important to stimulate students to complete science programs at the level of higher education.
- ''Inbound mobility rate''. In addition, the "middle" countries should attract more foreign students.

Note that the second-level relations are not harmful for countries that they do not explicitly address; rather, they are just not as important. 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; thus, they are not discussed further.

Compared to the relations from the year 2001, in 2010, we observed only a couple of changes. While the level of enrolment remained the most important measure and the ability to attract foreign students remained as important as it had before, stimulating students to study abroad did not appear important. At the second level, the single relation that was no longer important was stimulating students to complete science programs at the level of higher education.

Several of the presented relations have been recognized within the related work, showing the ability of the HMDM method to find important relations in the domain. The importance of the level of participation in higher education was recognized by Keller [30]. Furthermore, the importance of the investment in education was acknowledged by Gylfason [23], and the importance of participation in science-oriented higher education programs was noted by Varsakelis [63]. With the help of the HMDM method, we further discovered that a higher mobility of students is also very important for improved economic welfare. This relation was not directly discussed in the related work [23,30,63]. In contrast to the related work, our method provides a classification scheme that not only differentiates credible relations from less-credible relations, but also indicates how credible the discovered relations are and presents the relations in a human-readable form.

4.4. Analyses with the R&D data

Five relations emerged as credible. The ''GERD per capita (in PPP$)'' || ''GERD as % of GDP'' || ''GERD as % of GNI'' relation belongs to the first level of credibility, indicating that the most important factor for improving welfare is increasing the level of investment in R&D. GERD stands for Gross Domestic Expenditure on R&D, denoting the expenditure on R&D performed in a national territory during a given year [50]. PPP$ stands for purchasing power parity in American dollars. A statement of GERD in PPP$ allows for fair comparisons between countries. The importance of GERD has been acknowledged in economic literature. GERD is generally used as a control relation [63] to examine how successful a proposed method is at detecting important relations within a domain. The second level of credibility contains four relations. The first relation indicates that, to improve welfare, it is important to increase the number of employees in the R&D sector: ''Researchers per million inhabitants (head count)'' || ''Researchers per million inhabitants (full-time equivalent)'' and ''R&D personnel per million inhabitants (head count)'' || ''R&D personnel per million inhabitants (full-time equivalent)''. The second relation indicates that, for "low" countries, it is important to intensify work on innovations with the help of foreign experts: ''Application for patents (non-residents)''. The third relation indicates that the work on innovations should be intensified, as reflected in a higher number of patents: ''Applications for patents per million inhabitants'' and ''Grants of patents per million inhabitants''. The fourth relation indicates that the business enterprise sector should be the key leader in R&D activities: ''Sector employing the most researchers'' || ''Sector employing the most R&D personnel'' || ''Sector investing the most in R&D'' || ''Source of funds for R&D – Business enterprise (%)''. Finally, although there is some evidence that the amount of exported goods and services obtained as a result of intensive R&D activities is important for a country's welfare, the direction of this relation remains unexplained. The literature supports these conclusions (for example, Furman et al. [20]).

Relations for 2010 did not change considerably compared to those of 2001. Additional evidence was observed to indicate that having a government sector as the key leader in R&D activities negatively influences a country's welfare. ''Source of funds for R&D – Government (%)'' emerged as a relation of the second level of credibility.

The most important first-level credibility relations are the same for both years, with an exception being student mobility in the higher education domain.

5. Evaluation

The procedures in HMDM that are based on attribute selection mechanisms are similar to those in WAS, since they all construct multiple attribute sets and models to obtain the best results. Therefore, we begin the evaluation by drawing a line between the two methods. We then present an experimental setup, results of comparisons and a discussion of the results. This is followed by a user study, which was conducted to understand whether humans accept credible models as better than automatically constructed models. Finally, the evaluation is concluded by drawing a line between the current study and the authors' previous work.

5.1. HMDM vs. WAS in attribute selection and overall

The goal of WAS is to find the optimal subset of attributes that will result in the highest-quality possible prediction model that can be constructed with a specific DM method from a specific data set. The optimal subset does not need to contain all attributes relevant to the analysis. An attribute's relevance is defined as the systematic variation of the attribute's values

with category membership [21]. Kohavi and John [33] showed that predictors may benefit from the omission of relevant attributes and the addition of irrelevant attributes. For example, John [28] examined cases in which adding relevant attributes to the data set results in decision-tree models having lower quality.

Unlike WAS, HMDM emphasizes domain analysis and an understanding provided by humans. WAS optimizes for a formal maximum quality only, and, as a consequence, semantically irrelevant attributes and relations may exist in the constructed model. In contrast, HMDM optimizes for both meaning and quality by: (a) discovering as many relevant attributes as possible by constructing relations and models from them; (b) supporting the human while assessing the credibility of the relations and models; and (c) supporting the human while extracting credible relations and models. In addition, HMDM helps the human by providing insights during the process, while WAS keeps information hidden. Finally, the output of WAS is one model, while the output of HMDM consists of multiple relations and models, as well as human conclusions obtained through the analysis. It should be noted, however, that WAS is several times faster than HMDM and requires no human intervention.

In the following sections, we will compare HMDM and WAS in domain analysis tasks using two real-life domains.

5.2. Experimental setup

Two groups of experiments were conducted to compare HMDM and WAS.

5.2.1. Comparisons on a standard attribute set

For specific data D, a DM method I and a range of I's parameters P, HMDM constructs multiple models with I by changing the parameters and attributes in the prescribed manner (Section 3.3) to find credible relations and models in D that bear meaning in real life. A majority-vote ensemble is composed of the resulting set of credible models, which are obtained from a single iteration of HMDM on D. The ensemble is used as a representative of HMDM's capabilities.

Similarly, WAS uses I to construct multiple models by changing attributes in a prescribed manner, but with the goal of finding the attribute subset that will result in the highest quality model constructed from D. In this process, WAS does not consider whether the constructed models bear meaning in real life, and they often do not. The standard version of WAS also does not consider parameters P.

HMDM is compared with three versions of WAS: standard version (WAS-S), our modified version (WAS-MOD) and our modified version that outputs a majority-vote ensemble (WAS-ENS). The WAS-S algorithm begins with an empty set of attributes and adds new attributes as long as the addition of new attributes increases the quality. It is a Weka algorithm implemented as WrapperSubsetEval evaluator. We combined the evaluator with the GreedyStepwise search method, the same search mechanism used by HMDM in the ATTACH_ATTRIBUTES and DELETE_ATTRIBUTES procedures. Apart from setting I as a DM method, we use the default values for I's parameters and for all of the WAS-S algorithm's other parameters.

Modifications are introduced to provide fair many-to-many comparisons between HMDM and WAS, thus enabling both algorithms to search through approximately the same sized search space. More specifically, the parameter selection step is added as an internal step to WAS-MOD. The step invokes INITIAL_DM, which constructs all possible models within P, for each attribute subset evaluated by WAS-MOD. The quality of the highest quality model is considered. Finally, when WAS returns the optimal attribute subset, INITIAL_DM is employed. To represent the WAS-MOD's capabilities, the highest quality model from the set is considered. To represent the WAS-ENS's capabilities, all of the constructed models are included in the majority-vote ensemble.

5.2.2. Comparisons on constructed attributes

We repeated the experiments with the standard attribute set, with two modifications. First, the attribute construction step was employed in HMDM (Step 2—Section 3.3) and, second, automatic attribute construction was employed on a data set inputted to WAS. In the latter case, for D, new attributes were constructed by executing sum, min and max functions on pairs of attributes. Only pairs were considered, in order to obtain a manageable attribute set. In total, 5310 new attributes were added to the HI-EDU$_{class}$ data set, resulting in the HI-EDU$_{class\text{-}const}$ data set, and 3384 new attributes were added to the R&D$_{class}$ data set, resulting in the R&D$_{class\text{-}const}$ data set.

The comparisons were made for the following setups: (a) D = HI-EDU, I = J48, P = MNIL: 2–15, REP: on/off; and (b) D = R&D, I = J48, P = MNIL: 2–15, REP: on/off. WAS used ACC as the measure of quality, which was estimated with a 5-fold cross-validation (seed 1). The results for 2001 were obtained by applying a 10-fold cross-validation to the 2001 data. The results for 2010 were obtained by constructing models from the 2001 data and testing them on the 2010 data.

5.3. Results and discussion

5.3.1. Comparisons on a standard attribute set

For the 2001 data, the results of quality-based comparisons between HMDM and WAS show that WAS-MOD outperforms HMDM for both HI-EDU and R&D data (Table 3). In the case of HI-EDU data, WAS-ENS also outperforms HMDM. In contrast, when tested on the 2010 data, HMDM outperforms all WAS versions in the case of HI-EDU data, while the results are comparable for R&D data for WAS-ENS only. In terms of meaning, WAS constructed partially meaningless trees (similar to those analyzed and demonstrated in Fig. 8) (See Table 4).

Table 3
Quality-based comparisons between WAS and HMDM on a standard attribute set for 2001 and 2010.The highest-quality result is marked in bold.

Data set	$HI\text{-}EDU_{class}$				$R\&D_{class}$			
Year	2001		2010		2001		2010	
Algorithm	ACC	Kappa	ACC	Kappa	ACC	Kappa	ACC	Kappa
WAS-S	64.07	0.4102	41.60	0.0814	56.89	0.2674	71.79	0.5297
WAS-MOD	74.85	0.5893	56.80	0.3265	**68.26**	**0.4643**	75.64	0.5921
WAS-ENS	**75.40**	**0.5896**	56.80	0.3265	61.65	0.3633	**76.92**	0.6162
HMDM	73.13	0.5682	**60.80**	**0.3813**	65.77	0.4209	**76.92**	**0.6262**

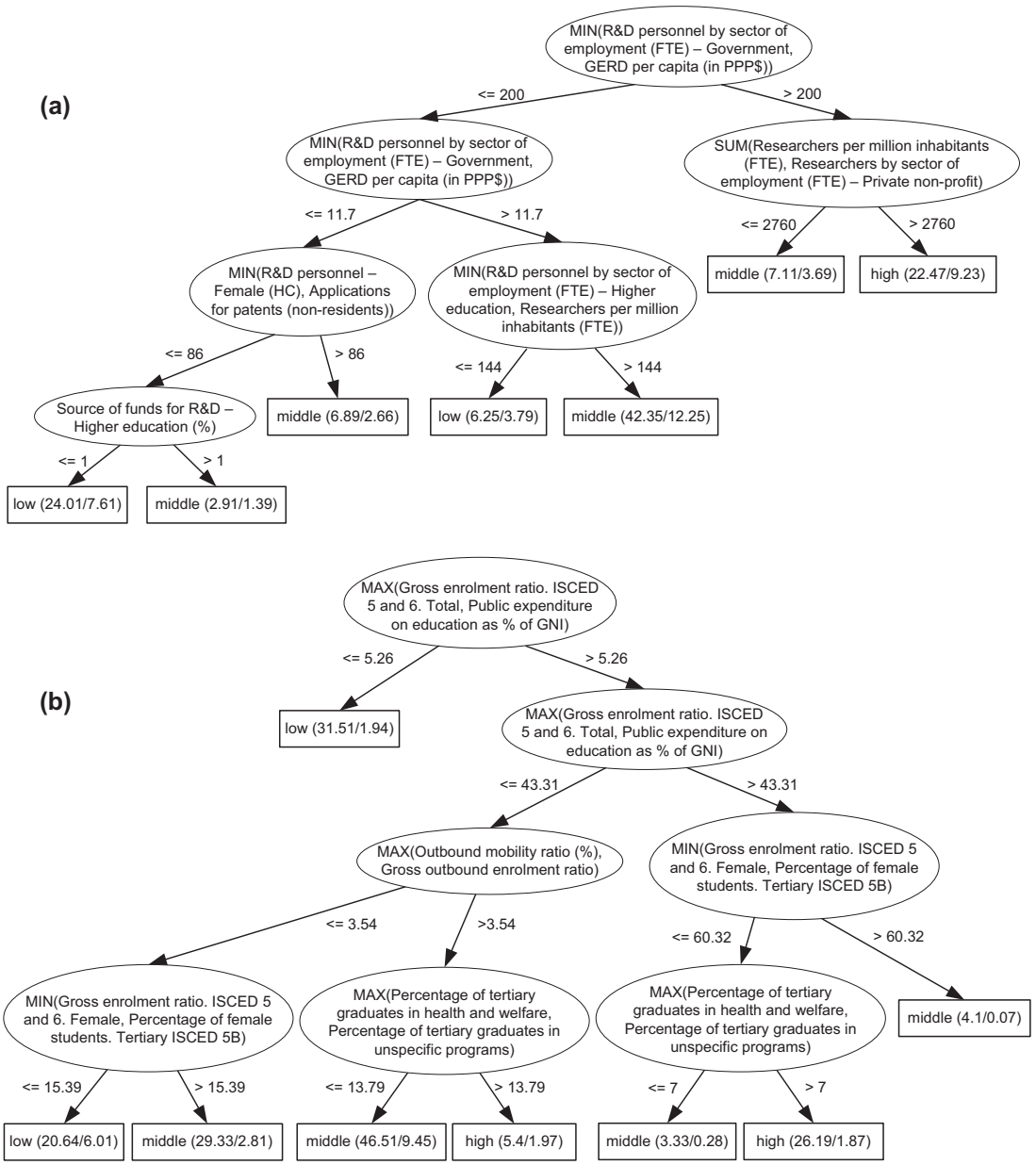

Fig. 8. Two decision trees constructed by WAS-MOD on (a) the $R\&D_{class\text{-}const}$ data set and (b) the $HI\text{-}EDU_{class\text{-}const}$ data set.

Table 4
Attribute sets and algorithm parameters used to construct the models in Table 3.

Data set	$HI\text{-}EDU_{class}$		$R\&D_{class}$	
Algorithm	Attributes	Parameters	Attributes	Parameters
WAS-S	(1) GER – Male; (2) Percentage of female students. Total tertiary; (3) Female tertiary graduates as percentage of all graduates in humanities and arts	Default	(1) Researchers per million inhabitants (HC); (2) GERD per capita (in PPP$); (3) Source of funds for R&D - Government (%); (4) Source of funds for R&D - Funds from abroad (%)	Default
WAS-MOD	(1) GER – Male; (2) GOER; (3) Tertiary students per 100,000 inhabitants	MNIL 15	(1) R&D personnel - Female (HC); (2) Total other supporting staff (HC); (3) Applications for patents (non-residents); (4) GERD per capita (in PPP$)	MNIL 15
WAS-ENS	The same as WAS-MOD.	MNIL: 2–15 REP: on/off	The same as WAS-MOD.	MNIL: 2–15 REP: on/off
HMDM (1)	(1) GER – Total; (2) GOER	MNIL 11	GERD per capita (in PPP$)	MNIL 2
HMDM (2)	(1) Outbound mobility ratio (%); (2) Tertiary students per 100,000 inhabitants	MNIL 14	(1) GERD per capita (in PPP$); (2) Applications for patents (non-residents)	MNIL 13
HMDM (3)	(1) Tertiary students per 100,000 inhabitants; (2) PE-GNI	MNIL 15	(1) GERD as % of GDP; (2) Applications for patents (non-residents)	MNIL 13 REP
HMDM (4)	(1) Tertiary students per 100,000 inhabitants; (2) Public expenditure on education as a % of GDP	MNIL 15	(1) GERD as % of GDP; (2) Researchers per million inhabitants (HC)	MNIL 14
HMDM (5)	(1) GER – Total; (2) PE-GNI	MNIL 11		
HMDM (6)	(1) GER – Total; (2) CE-GNI	MNIL 9		

5.3.2. Comparisons on constructed attributes

When new attributes were constructed, the quality generally increased for the 2001 data and decreased for the 2010 data for both WAS and HMDM with regard to HI-EDU and increased for both years with regard to R&D (Tables 3 and 5). On the constructed 2001 data (Table 5), WAS-MOD outperformed HMDM in terms of quality. WAS-ENS also outperformed HMDM for HI-EDU 2001 data. In contrast, HMDM outperformed all versions of WAS on the 2010 constructed data. Together, the results in Tables 3 and 5 indicate that HMDM performs equally well or better than WAS when learned on 2001 and tested on 2010 data. The 2010 results indicate that HMDM generalized better than WAS, which over-fitted to data in 2001.

When the models constructed by the three versions of WAS (Table 6) were examined for meaning, different types of meaningless relations were detected, including misleading relations and relations not interpretable in the context of the analysis.

For example, a misleading relation appeared in the left-under-top node in the tree in Fig. 8a, which was constructed by WAS-MOD. Examining several variations of this tree with HMDM reveals that the relation including the constructed attribute MIN (R&D personnel by sector of employment (FTE) – Government (R&D-PERS-GOV)), GERD per capita (in PPP$) is misleading because the splitting points for that relation in Fig. 8a are practically the same as those for the relations that only include the standard attribute GERD per capita (in PPP$) in the alternative trees. Indeed, the MIN of the two attributes for all (except the two) instances in the data set is the value of the GERD attribute. In addition, in most of the trees produced during the HMDM session, R&D-PERS-GOV proved irrelevant and rarely had a positive influence on a country's welfare. WAS, on the contrary, found R&D-PERS-GOV to have a positive effect twice in the tree in Fig. 8a.

An example of a relation that is not interpretable in the context of the analysis appeared twice in Fig. 8b (bottom of the tree). This relation includes the constructed attribute MAX (Percentage of tertiary graduates in health and welfare, Percentage of tertiary graduates in unspecific programs). If the goal of the analysis is to determine which segments of the higher education sector influence economic welfare, this relation would state that if the country increases the share of tertiary graduates in health and welfare or unspecified programs, a country's welfare will increase as well. Using HMDM, such relations are examined and clarified through the construction of additional attributes and models.

In summary, HMDM outperformed WAS when testing the models constructed from the 2001 data on the 2010 data, indicating that credible relations and models provide well-generalized domain descriptions. While, on the same year, WAS does better identify models of improved quality than HMDM, it is susceptible to the problems mentioned in the introduction due

Table 5
Quality-based comparisons between WAS and HMDM on constructed attributes for the years 2001 and 2010. The highest-quality result is marked in bold.

Data set	$HI\text{-}EDU_{class\text{-}const}$				$R\&D_{class\text{-}const}$			
Year	2001		2010		2001		2010	
Algorithm	ACC	Kappa	ACC	Kappa	ACC	Kappa	ACC	Kappa
WAS-S	68.86	0.5002	37.60	−0.0146	65.87	0.4223	42.31	0.0389
WAS-MOD	**82.63**	**0.7144**	45.60	0.0582	**71.26**	**0.5065**	71.79	0.5111
WAS-ENS	79.71	0.6715	46.40	0.1005	65.18	0.3942	66.67	0.4230
HMDM	74.93	0.5942	**52.00**	**0.2076**	68.13	0.4547	**83.33**	**0.7209**

Table 6
Attribute sets and algorithm parameters used to construct the models in Table 5.

Data set	$HI\text{-}EDU_{class\text{-}const}$(WAS); $HI\text{-}EDU_{class\text{-}mod}$(HMDM)		$R\&D_{class\text{-}const}$(WAS); $R\&D_{class\text{-}mod}$(HMDM)	
Algorithm	Attributes	Parameters	Attributes	Parameters
WAS-S	(1) MAX (GER – Total, PE-GNI); (2) MAX (GER – Total, CE-GNI); (3) SUM (GER – Male, Outbound mobility ratio (%)); (4) SUM (Distribution of students (%). ISCED Level 5B, CE-GNI); (5) SUM (Percentage of tertiary graduates in services, Educational expenditure by nature of spending as a % of total educational expenditure on public institutions. Tertiary. Total current expenditure)	Default	(1) MIN (Researchers per million inhabitants (FTE), Researchers per million inhabitants (HC)); (2) MIN (Grants of patents (residents), Source of funds for R&D – Higher education (%)); (3) SUM (GERD per capita (in PPP$), Source of funds for R&D – Not distributed funds (%))	Default
WAS-MOD	(1) MIN (GER – Total, Percentage of female graduates in tertiary education); (2) MAX (GER – Total, PE-GNI); (3) MIN (GER – Female, Percentage of female students. Tertiary ISCED 5B); (4) MAX (Outbound mobility ratio (%), GOER); (5) MAX (Percentage of tertiary graduates in health and welfare, Percentage of tertiary graduates in unspecified programs)	MNIL 3	(1) Source of funds for R&D – Higher education (%); (2) MIN (R&D personnel – Female (HC), Applications for patents (non-residents)); (3) MIN (R&D personnel by sector of employment (FTE) – Government, GERD per capita (in PPP$)); (4) MIN (R&D personnel by sector of employment (FTE) – Higher education, Researchers per million inhabitants (FTE)); (5) SUM (Researchers per million inhabitants (FTE), Researchers by sector of employment (FTE) – Private non-profit)	MNIL 2; REP
WAS-ENS	The same as WAS-MOD.	MNIL: 2–15, REP: on/off	The same as WAS-MOD.	MNIL: 2–15, REP: on/off
HMDM (1)	(1) GER – Total + GOER; (2) GOER; (3) GER – Total + PE-GNI	MNIL 13	(1) GERD per capita (in PPP$); (2) Applications for patents (non-residents) (%); (3) Grants of patents per million inhabitants	MNIL 10
HMDM (2)	(1) GER – Total + GOER; (2) GOER; (3) Percentage of tertiary graduates in science	MNIL 8	(1) GERD per capita (in PPP$); (2) Applications for patents (non-residents) (%); (3) Applications for patents (non-residents)	MNIL 10
HMDM (3)	(1) GER – Total + GOER; (2) GOER; (3) Inbound mobility rate	MNIL 8	(1) GERD per capita (in PPP$); (2) Applications for patents per million inhabitants	MNIL 13
HMDM (4)	(1) GER – Total + GOER; (2) GOER; (3) GER – Total	MNIL 12	(1) GERD as % of GNI; (2) Applications for patents per million inhabitants	MNIL 9
HMDM (5)	(1) GER – Total; (2) GOER	MNIL 11	(1) GERD as % of GNI; (2) Researchers per million inhabitants (HC)	MNIL 7
HMDM (6)	(1) GOER; (2) GER – Total + CE-GNI	MNIL 11 REP	(1) GERD as % of GNI; (2) Researchers per million inhabitants (FTE)	MNIL 8
HMDM (7)	(1) GER – Total; (2) GOER; (3) GER – Total + PE-GNI	MNIL 12 REP	(1) GERD as % of GDP; (2) Applications for patents per million inhabitants	MNIL 11
HMDM (8)	(1) Outbound mobility ratio (%); (2) Tertiary students per 100,000 inhabitants; (3) PE-GNI	MNIL 11	(1) GERD per capita (in PPP$); (2) Grants of patents per million inhabitants	MNIL 11
HMDM (9)	(1) GER – Total + GOER; (2) PE-GNI	MNIL 5		

to its reliance on objective criterion only and not on incorporating human knowledge and understanding. However, WAS can be an important addition to the HMDM process, since it constructs models differently from other DM methods. In practical terms, the use of WAS is recommended, as is the use of other methods that contribute alternative viewpoints.

5.3.3. Discussion on complexity

Since the "common sense and informal domain knowledge" part of the meaning criterion cannot be formalized, the human must be included in the DM loop when searching for credible relations and models. A side effect of introducing the human input into HMDM is a reduction in the complexity of HMDM's computer part.

This reduction may be achieved in the attribute construction and domain searching steps (Steps 2 and 6 in Section 3.3). In HMDM, the human constructs interesting new attributes by combining one or several existing attributes. For example, the human constructed the attribute "Sector employing the most R&D personnel" by computing the maximum value among the four attributes representing the distribution of personnel over the four sectors. The relations containing the attribute were interpretable in the context of the analysis and were proven to be the second level of credibility relations (Section 4.4). While the human constructed and tested a single interesting attribute, without the human input $\binom{n}{4}$ new attributes would have to be constructed and tested in order to determine whether the "sector" relation was credible (assuming that only combinations of the four attributes were interesting). In the case of the R&D and HI-EDU data, this means that 194,580 and 487,635 new attributes would have to be constructed and tested, respectively.

In the domain searching step, the worst-case complexity of automatic DELETE_ATTRIBUTES and ATTACH_ATTRIBUTES procedures is $n!$ In practice, the complexity of the DELETE_ATTRIBUTES procedure is controlled by deleting only the attributes that appear in

models and by the greedy nature of the procedure. Similarly, the human uses the ATTACH_ATTRIBUTES procedure with the goal of reexamining the credibility of the relations that emerge in the delete attributes step, which reduces the number of attributes n. The domain expert focuses on specific relations, thus requiring the construction of fewer models to test the hypotheses.

The time complexity of the complete interactive process was assessed on 20 data sets describing web genres [65], each of which was composed of 1539 examples and 500 attributes. In this study, the novice needed two or three working days per data set to extract credible relations and models. The exceptions comprised five data sets, for which the first cycle of analysis showed that the selected attributes could not represent the genre.

5.3.4. Discussion on subjectivity

HMDM is a systematic method, but it is also subjective because it involves an interplay between computer-supported DM methods and human judgment. For this reason, the results of using the method are subjective and may differ between users. Moreover, different users can apply different strategies for using the method. In our experience, the first level of credibility relations and models could be found equally easy by both novices and experts. The differences typically appeared in relations of the second and third levels of credibility. In our experiments, the first level of credibility relations were always supported by strong evidence that could not be overlooked. The second and the third level of credibility relations sometimes differed between users, but the differences were not major. However, a novice HMDM user might introduce two types of mistakes due to subjectivity. First, he/she might not notice a potential candidate and stop the search too soon. Second, he/she might prefer a quantitatively worse hypothesis over a better one, in which case he/she would have to ignore non-confirming evidence. In summary, subjectivity can change the order of the relations, but only to a limited degree, since the relations are grouped into three categories.

5.4. User study

With the help of a user study, we aimed to determine: (Q1) whether humans recognize the less-credible relations in a single decision-tree model constructed by an automatic method and (Q2) whether they accept credible trees as better than automatically constructed trees.

We extracted a set of trees from the experiment in the R&D domain (Section 4.4) and organized them in the form of a paper-based questionnaire. In total, there were 22 participants in the study, all of whom had prior knowledge of the decision trees.

The experiment was conducted one participant at a time, with a facilitator interacting with each participant. Since the participants were not domain experts and, thus, were not familiar with all the attributes, the facilitator helped by answering technical questions. The facilitator, however, did not influence the participant's choices or indicate which method was used to construct the presented trees. The participants were encouraged to express any comments they had regarding the task. The experiment was performed, page-by-page, in one pass, with no moving backward or looking forward in the process. The data set, the questionnaire and the table with all the answers are available at http://dis.ijs.si/Vedrana/user-study.htm.

The questionnaire is composed of two parts. Altogether, it comprises four trees: one presented in Fig. 2 and three in Fig. 9a–c. Similar to the trees in Fig. 9, the tree in Fig. 2 had additional information attached regarding its quality: ACC 64.67%; CCPE 0.4113; Kappa 0.4047; CPX 11. The first part of the questionnaire corresponds to Q1 and Fig. 2, and this was followed by questions intended to determine whether the participants found the presented tree reasonable. We used the term "reasonable" without any additional explanation to observe the criteria on which the participants based 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. With respect to Q1, we can see that 55% of participants accepted the automatically constructed tree and all the relations contained within the tree (four "yes" answers).

The participants sometimes commented that the structure of the tree was strange, for example, due to branches in the right subtree with no included examples or the presence of a N/A branch that did not bear important information; however, they considered the tree to be semantically meaningful, which prevailed in the positive decision. A total of 23% of the participants correctly stated that "the level of investment in R&D" is important and that the two subtrees were unreasonable ("yes no no" answers to questions 2 through 4). About 18% of the participants correctly stated that "the level of investment in R&D" is important and marked only one of the two subtrees as unreasonable ("yes yes no" or "yes no yes" answers to questions 2 through 4). Finally, only 4% of the participants marked the tree and all of its relations as unreasonable. Considering that only 23% of the participants correctly detected credible relations from a single tree, this illustrates the need for analysis tools such as HMDM to examine relations for credibility.

The second part corresponds to Q2 and reveals three credible trees, which are followed by questions that sought to determine whether the participants accepted the three credible trees as more credible than the tree constructed by the automatic method and on which criteria they based their decision. The credible tree in Fig. 9a was denoted as the second tree (since the first was the tree in Fig. 2), the tree in Fig. 9b as the third, and the tree in Fig. 9c as the fourth. Each tree was supported by the

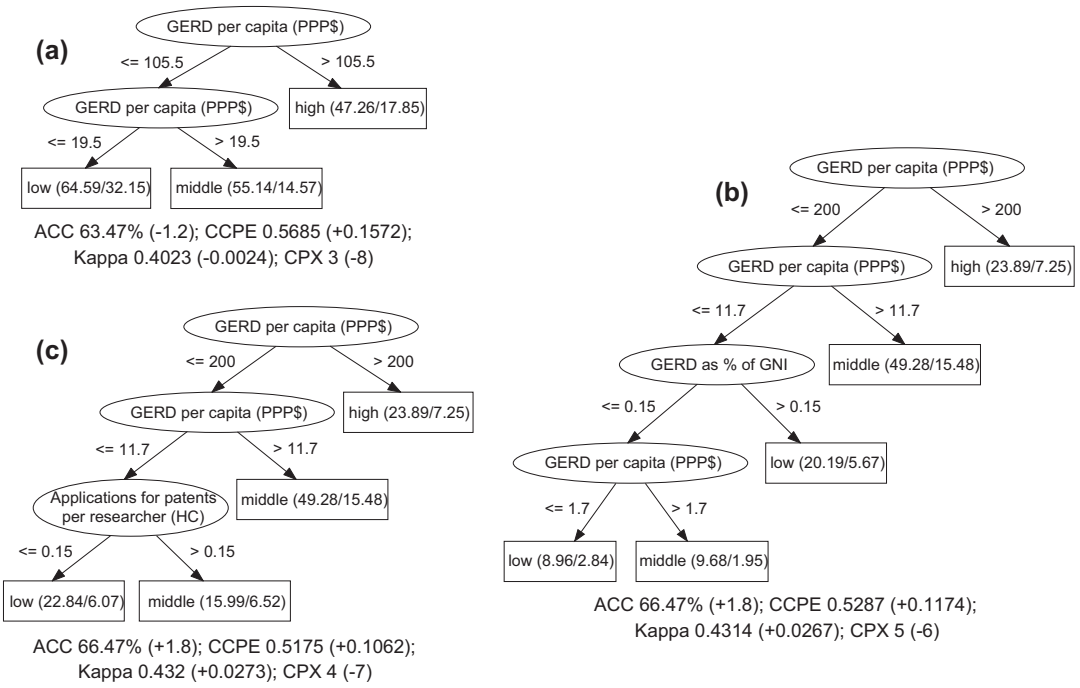

Fig. 9. Three credible trees presented within the questionnaire.

Table 7
The results of the user study.

	Questions						
Participants	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
Summary	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

quality measures and the differences in quality from the first tree. Each participant would obtain the same view of the tree by clicking on the node of the attached attributes graph from which the tree was extracted; however, the participants did not have access to the whole program—just to one sheet of paper at a time.

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, complexity (CPX), other measures: CCPE and/or Kappa (C/K), content of the tree (cont.)); (7) What was the sequence of trees that persuaded you the most that the GERD attributes are the most important for the welfare of a country? (For example, 2, 3, 1, 4 meant that 2 was the most persuasive and 4 the least persuasive.)

With respect to Q2, all the participants stated that at least one of the three additional trees sounded more reasonable than the first one (question 5), indicating that the participants generally accepted the credible trees as better than the tree constructed by the automatic method. Approximately 77% of the participants selected the tree in Fig. 9c, 73% selected the tree in Fig. 9a, and 27% selected the tree in Fig. 9b as more credible than the first tree. The answers to question 5 were supported by the answers to question 7, in which 64% of the participants ranked all the credible trees higher than the first tree.

The most-selected criterion for choosing the credible trees as more reasonable than the first tree was 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, and 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 measures (ACC, CCPE/Kappa) played a part in their decision. The participants frequently stated that the differences in quality were small, which was the most probable reason for their not selecting quality as the relevant criterion. However, such behavior is desirable in the case of the HMDM method, since the system pre-selects trees based on quality, and it is up to the human to make the final decisions based primarily on the content. Additionally, the participants saw only a couple of best trees of similar estimates—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, as a way of providing additional information when comparing similar models.

In conclusion, the participants changed their opinions in favor of the credible trees based primarily on the better-estimated and more-understandable content of these trees, even though they were able to see only the three best trees. It is assumed that the participants would have been even more convinced had they had additional access to the program online.

5.5. Comparisons with the previous work of authors

The work presented in this paper is an extension of the preliminary work published in Vidulin and Gams [64]. Table 8 explains the changes. In addition, there are four new contributions. First, one of the measures applied to the decision trees is the statistical measure CCPE. This measure was originally designed for classification rules, but we have adjusted it here to work on decision trees. Second, in contrast to the unfair one-to-many quality-based comparison in [64], the current study provides an improved many-to-many quality-based comparison. Third, the improved method is evaluated with a user study. Fourth, an extensive comparison with WAS was performed on data from two years: 2001 and 2010.

6. Conclusions and discussion

In this paper, 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

Table 8
Improvements in the present study compared to the preliminary work [64].

Improvement	Preliminary work	Current work
Refined relation-extraction scheme	The scheme defines three levels of credibility	The scheme defines two additional types of relations: redundancy and combination. The heuristic function, called the interaction, is added to provide additional evidence indicating a relation's type
Additional quality measures to facilitate the extraction of credible relations	Quality-based comparisons are based on a single measure: ACC for decision and CC for regression trees	Quality-based comparisons are based on three measures (ACC, CCPE and Kappa) for decision trees and two measures (CC and RAA) for regression trees
Refined interactive explanation structures: deleted and attached attributes graphs	For each model, the structures include only one quality assessment and no means for quality-based comparisons between the models	The structures include three quality assessments. The models are easily compared by observing the q_A measure and the differences in individual assessments that follow each measure
Improved capability of the HMDM algorithm to find second-level relations	Able to show first-level relations and indicate other relations	This version explicitly indicates the second-level relations
Enabled automatic classification from human-stored credible models	HMDM is able to store the credible models	HMDM stores and constructs a majority-vote ensemble of credible models for automatic classification purposes

human insight and understanding. The implemented interactive system constructs multiple models and arranges them in deleted and attached attributes graphs. The human observes the constructed models and, with the help of the relation-extraction scheme, extracts credible relations and models, which are more meaningful and of higher quality for new data than the models constructed with automatic DM methods.

The HMDM method is designed to be an interactive rather than an automatic method for two reasons. First, HMDM is designed to support a human in the domain analysis process by enabling him/her to interactively explore and learn about the domain of interest. HMDM supports novices, who are exploring a domain for the first time by examining the relations that emerge and then using common knowledge or studying the literature to explain the obtained relations. HMDM also supports experts by allowing them to test hypotheses based on previous knowledge about a domain. Second, it is difficult to formalize a domain knowledge or common sense. Therefore, the HMDM method is designed to combine the best of both worlds: formal measures of quality and informal knowledge provided by a human.

The application of HMDM was demonstrated on two complex domains in order to answer the question: Which segments of the higher education and R&D sectors influence the economic welfare of a country? The results of the analysis showed that the HMDM method is capable of finding important relations in a domain: With HMDM, it was established that, for better welfare, it is important to increase the level of investment in R&D; in economic literature, this relation is used as a control relation to examine how successful the proposed method is at detecting important relations in the domain. In addition, with the help of the HMDM, we also discovered a relation that was not directly discussed in the literature: that a higher ability to attract foreign students is important for better economic welfare.

The HMDM method was further verified through quality-based comparisons with three versions of WAS. The results indicate that HMDM has better generalization capabilities than WAS: HMDM outperformed WAS on the test data collected for a year that was different than the one from which the models were constructed. The results showed that credible relations constructed by HMDM remain valid, even after a nine-year period. Further analysis indicated that differences in generalization capabilities emerge mostly due to the appearance of meaningless relations in WAS models, which were manifested, for example, in the form of: misleading relations and relations not interpretable in the context of analysis. In addition, one of the benefits of HMDM is that the knowledge that the human gathers during the process itself is as valuable as the knowledge gained from the results. Note that the human mental model is upgraded during the process without any formal or semantic limitations.

The final evaluation step included a user study, which showed that all of the 22 participants accepted one or several of the credible decision trees as better than the automatically constructed trees.

A question remains open regarding whether additional relations would be extracted by different humans performing HMDM. To verify that the method is stable and that the results do not vary significantly depending on the particular human performing the HMDM, the data and results have been made available on the Internet (http://dis.ijs.si/Vedrana/economic-analysis.htm). The HMDM program is available at http://dis.ijs.si/Vedrana/HMDM.htm.

Another debate is open regarding which type of relation was indeed observed: X implies Y or Y implies X. Does more investment in the R&D sector actually cause countries to progress faster, or is spending more on R&D just a side effect of developed countries? Although the 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. However, to evaluate this relation in a quantitative way, other methodologies are more appropriate than HMDM.

In general, human ingenuity is critical to the acceptance or rejection of any conclusion supported by statistics or any other formal method. By observing, not only one model in one DM setup, but thousands of models in the process of creation, and by providing an interactive tool to verify the hypotheses, thus enabling the human mind to integrate conclusions from thousands of constructed transparent models, we show that the summarized relations indeed emerge as credible.

As part of future work, five improvements of HMDM seem interesting. First, in this paper, we tested the HMDM in combination with decision and regression trees; however, we assume that HMDM is applicable to other supervised DM methods that produce models in a human-understandable form. As part of future work, we plan to test HMDM in combination with other DM methods. Second, the flexible quality criterion for ranking the models within the deleted and attached attributes graphs could be implemented. Flexible means that the human can attribute different weights to the selected quality measures and, in this manner, tune the algorithm to give higher ranks to those models he/she considers more credible. Third, the weights within the quality-based ranking criterion may be learned from the models marked as credible by the human. In contrast, learning the meaning-based criterion from the credible models is problematic, since it is hard to formalize domain knowledge and common sense. Fourth, a tool for selecting combinations of attributes that do not make sense to the human will be added in Step 4 of the method. The human's choices will be translated into constraints, which will be used to eliminate the models with uninteresting combinations from further search. Finally, we intend to improve the visualization of the deleted and attached attributes graphs.

Acknowledgements

The authors would like to thank the anonymous reviewers for their valuable comments and suggestions to improve the quality of the paper. We are also grateful to Sanja Kovač, who contributed to the experimental evaluation, and Jure Grabnar, who contributed to the HMDM program.

References

[1] R. Aliev, W. Pedrycz, B. Fazlollahi, O.H. Huseynov, A.V. Alizadeh, B.G. Guirimov, Fuzzy logic-based generalized decision theory with imperfect information, Inform. Sci. 189 (2012) 18–42.
[2] S. Badr, A. Bargiela, Case study of inaccuracies in the granulation of decision trees, Soft. Comput. 15 (2011) 1129–1136.
[3] B. Becker, R. Kohavi, D. Sommerfield, Visualizing the simple bayesian classifier, in: U. Fayyad, G. Grinstein, A. Wierse (Eds.), Information Visualization in Data Mining and Knowledge Discovery, Morgan Kaufman, San Francisco, CA, 2001, pp. 237–249.
[4] J. Black, N. Hashimzade, G. Myles, A Dictionary of Economics, Oxford University Press, New York, 2009.
[5] M. Bohanec, I. Bratko, Trading accuracy for simplicity in decision trees, Mach. Learn. 15 (1994) 223–250.
[6] L. Breiman, Random forests, Mach. Learn. 45 (2001) 5–32.
[7] L. Cao, P.S. Yu, C. Zhang, Y. Zhao, Domain Driven Data Mining, Springer, 2010.
[8] B. Chandra, P. Varghese, Moving towards efficient decision tree construction, Inform. Sci. 179 (2009) 1059–1069.
[9] Y. Chen, Y. Yao, A multiview approach for intelligent data analysis based on data operators, Inform. Sci. 178 (2008) 1–20.
[10] J. Cohen, A coefficient of agreement for nominal scales, Educ. Psychol. Meas. 20 (1960) 37–46.
[11] D.A. Cohn, Z. Ghahramani, M.I. Jordan, Active learning with statistical models, J. Artif. Intell. Res. 4 (1996) 129–145.
[12] Computational User Experiences Microsoft Research Group, Interactive Machine Learning, <http://research.microsoft.com/en-us/um/redmond/groups/cue/IML>.
[13] E. Costello, L. McGinty, Supporting user interaction in flavor sampling trials, in: Proc. of Workshop on Intelligence and Interaction, IJCAI 2009, Pasadena, CA, 2009.
[14] M.W. Craven, Extracting Comprehensible Models from Trained Neural Networks, PhD thesis, School of Computer Science, University of Wisconsin, Madison, WI, 1996.
[15] A. Culotta, T. Kristjansson, A. McCallum, P. Viola, Corrective feedback and persistent learning for information extraction, Artif. Intell. 170 (2006) 1101–1122.
[16] J. Demšar, B. Zupan, G. Leban, T. Curk, Orange: from experimental machine learning to interactive data mining, in: Proc. of the 8th Eur. Conf. on Principles and Practice of Knowledge Discovery in Databases, Pisa, IT, 2004, pp. 537–539.
[17] K. Dey, S. Rosenthal, M. Veloso, Using interaction to improve intelligence: how intelligent systems should ask users for input, in: Proc. of Workshop on Intelligence and Interaction, IJCAI 2009, Pasadena, CA, 2009.
[18] J.A. Fails, D.R. Olsen Jr., Interactive machine learning, in: Proc. of the 8th Int. Conf. on Intelligent User Interfaces, Miami, FL, 2003, pp. 39–45.
[19] A.J. Feelders, Prior knowledge in economic applications of data mining, in: Principles of Data Mining and Knowledge Discovery, LNCS 1910, Springer, 2000, pp. 395–400.
[20] J.L. Furman, M.E. Porter, S. Stern, The determinants of national innovative capacity, Res. Policy 31 (2002) 899–933.
[21] J.H. Gennari, P. Langley, D. Fisher, Models of incremental concept formation, Artif. Intell. 40 (1989) 11–61.
[22] I. Guyon, A. Elisseeff, An introduction to variable and feature selection, J. Mach. Learn. Res. 3 (2003) 1157–1182.
[23] T. Gylfason, Natural resources, education, and economic development, Eur. Econ. Rev. 45 (2001) 847–859.
[24] E.R. Hruschka, N.F.F. Ebecken, Extracting rules from multilayer perceptrons in classification problems: a clustering-based approach, Neurocomputing 70 (2006) 384–397.
[25] Y. Huang, T.M. Mitchell, Text clustering with extended user feedback, in: Proc. of the Int. SIGIR Conf. on R&D in Information Retrieval, SIGIR, Seattle, WA, 2006, pp. 413–420.
[26] A. Jakulin, Machine Learning Based on Attribute Interaction, PhD thesis, Faculty of Computer and Information Science, University of Ljubljana, Sežana, Slovenia, 2005.
[27] D. Jensen, P.R. Cohen, Multiple comparisons in induction algorithms, Mach. Learn. 38 (2000) 309–338.
[28] G.H. John, Enhancements to the Data Mining Process, PhD thesis, Computer Science Department, Stanford University, CA, 1997.
[29] A. Kapoor, B. Lee, D. Tan, E. Horvitz, Performance and preferences: interactive refinement of machine learning procedures, in: Proceedings of the 26th AAAI Conference on Artificial Intelligence, 2012, pp. 1578–1584.
[30] K.R.I. Keller, Investment in primary, secondary, and higher education and the effects on economic growth, Contemp. Econ. Policy 24 (2006) 18–34.
[31] S. Kelling, C. Lagoze, W.-K. Wong, J. Yu, T. Damoulas, J. Gerbracht, D. Fink, C. Gomes, eBird: a human/computer learning network to improve biodiversity conservation and research, AI Mag. 34 (2013) 10–20.
[32] R. Kohavi, A study of cross-validation and bootstrap for accuracy estimation and model selection, in: Proc. of Int. Joint Conf. on Artif. Intell., IJCAI, Montréal, Québec, Canada, 1995, pp. 1137–1145.
[33] R. Kohavi, G.H. John, Wrappers for feature subset selection, Artif. Intell. 97 (1997) 273–324.
[34] I. Kononenko, Inductive and bayesian learning in medical diagnosis, Appl. Artif. Intell. 7 (1993) 317–337.
[35] T. Kulesza, W-K. Wong, S. Stumpf, S. Perona, R. White, M.M. Burnett, I. Oberst, A.J. Ko, Fixing the program my computer learned: barriers for end users, challenges for machine, in: Proc. of Conf. on Intelligent User Interfaces, IUI 2009, Sanibel Island, FL, 2009.
[36] P. Kuntz, F. Guillet, R. Lehn, H. Briand, A user-driven process for mining association rules, in: D. Zighed, J. Komorowski, J. Zytkow (Eds.), Principles of Data Mining and Knowledge Discovery, Springer, Berlin, 2000, pp. 319–339.
[37] C.H. Liu, R. Matthews, Vygotsky's philosophy: constructivism and its criticisms examined, Int. Educat. J. 6 (2005) 386–399.
[38] H. Luan, Y.-T. Zheng, M. Wang, T.-S. Chua, VisionGo: towards video retrieval with joint exploration of human and computer, Inform. Sci. 181 (2011) 4197–4213.
[39] D.J.C. MacKay, Information-based objective functions for active data selection, Neural Comput. 4 (1992) 590–604.
[40] Y. Maeno, Y. Ohsawa, Human–computer interactive annealing for discovering invisible dark events, IEEE Trans. Indus. Electron. 54 (2007) 1184–1192.
[41] K.V. Mardia, J.T. Kent, J.M. Bibby, Multivariate Analysis, Academic Press, London, 1979.
[42] D. Martens, B. Baesens, Building acceptable classification models, in: Data Mining: Annals of Information Systems, Springer, 2010, pp. 53–74.
[43] D. Martens, J. Vanthienen, W. Verbeke, B. Baesens, Performance of classification models from a user perspective, Decis. Support Syst. 51 (2011) 782–793.
[44] P. Melville, S.M. Yang, M. Saar-Tsechansky, R. Mooney, Active learning for probability estimation using Jensen–Shannon divergence, in: Proc. of the 16th Eur. Conf. on Machine Learning, ECML 2005, Porto, Portugal, 2005, pp. 268–279.
[45] V. Mirchevska, Behavior Modeling by Combining Machine Learning and Domain Knowledge, PhD thesis, Jožef Stefan International Postgraduate School, Ljubljana, Slovenia, 2013.
[46] M. Možina, J. Demšar, M. Kattan, B. Zupan, Nomograms for visualization of naive bayesian classifier, in: Proc. of 8th Eur. Conf. on Principles and Practice of Knowledge Discovery in Databases, Pisa, IT, 2004, pp. 337–348.
[47] M. Možina, J. Demšar, J. Žabkar, I. Bratko, Why is rule learning optimistic and how to correct it, in: J. Fürnkranz, T. Scheffer, M. Spiliopoulou (Eds.), Machine Learning: ECML 2006, Springer, Berlin, 2006, pp. 330–340.
[48] F. Nasoz, C.L. Lisetti, A.V. Vasilakos, Affectively intelligent and adaptive car interfaces, Inform. Sci. 180 (2010) 3817–3836.
[49] T.D. Nguyen, T.B. Ho, H. Shimodaira, A visualization tool for interactive learning of large decision trees, in: Proc. of 12th IEEE Int. Conf. on Tools with Artif. Intell., ICTAI 2000, Vancouver, BA, Canada, 2000, pp. 28–35.
[50] OECD, Frascati Manual: Proposed Standard Practice for Surveys on Research and Experimental Development, OECD, Paris, 2002.
[51] K.M. Osei-Bryson, Evaluation of decision trees: a multi-criteria approach, Comput. Oper. Res. 31 (2004) 1933–1945.
[52] R.K. Pearson, Mining Imperfect Data, SIAM, 2005.
[53] R. Porter, J. Theiler, D. Hush, Interactive machine learning in data exploitation, Comput. Sci. Eng. (2013). preprint.

[54] Proceedings of the IUI 2013 Workshop on Interactive Machine Learning, IUI Workshop on Interactive Machine Learning, Located at 18th International Conference on Intelligent User Interfaces, March 19–22, Santa Monica, CA, USA, ACM, 3, 2013.
[55] J.R. Quinlan, Learning with continuous classes, in: Proc. of 5th Australian Joint Conf. on AI, Singapore, 1992, pp. 343–348.
[56] J.R. Quinlan, C.45: Programs for Machine Learning, Morgan Kaufman, San Mateo, CA, 1993.
[57] P.R. Rijnbeek, J.A. Kors, Finding a short and accurate decision rule in disjunctive normal form by exhaustive search, Mach. Learn. 80 (2010) 33–62.
[58] M. Shilman, D. Tan, P. Simard, CueTIP: a mixed-initiative interface for correcting handwriting errors, in: Proc. of the Sym. on User Interface Software and Technology, UIST 2006, Montreux, Switzerland, 2006, pp. 323–332.
[59] S. Stumpf, V. Rajaram, L. Li, W. Wong, M. Burnett, T. Dietterich, E. Sullivan, J. Herlocker, Interacting meaningfully with machine learning systems: three experiments, Int. J. Hum.–Comput. St. 67 (2009) 639–662.
[60] E. Štrumbelj, I. Kononenko, M. Robnik Šikonja, Explaining instance classifications with interactions of subsets of feature values, Data Knowl. Eng. 68 (2009) 886–904.
[61] S. Tong, D. Koller, Support vector machine active learning with applications to text classification, J. Mach. Learn. Res. 2 (2002) 45–66.
[62] G. Towell, J.W. Shavlik, Extracting refined rules from knowledge-based neural networks, Mach. Learn. 13 (1993) 71–101.
[63] N.C. Varsakelis, Education, political institutions and innovative activity: a cross-country empirical investigation, Res. Policy 35 (2006) 1083–1090.
[64] V. Vidulin, M. Gams, Impact of high-level knowledge on economic welfare through interactive data mining, Appl. Artif. Intell. 25 (2011) 267–291.
[65] V. Vidulin, Searching for Credible Relations in Machine Learning, PhD thesis, Jožef Stefan International Postgraduate School, Ljubljana, Slovenia, 2012.
[66] X. Wang, C. Dong, Improving generalization of fuzzy if–then rules by maximizing fuzzy entropy, IEEE Trans. Fuzzy Syst. 17 (2009) 556–567.
[67] X. Wang, L. Dong, J. Yan, Maximum ambiguity-based sample selection in fuzzy decision tree induction, IEEE Trans. Knowl. Data Eng. 24 (2012) 1491–1505.
[68] M. Ware, E. Frank, G. Holmes, M. Hall, I.H. Witten, Interactive machine learning: letting users build classifiers, Int. J. Hum.–Comput. St. 55 (2001) 281–292.
[69] I.H. Witten, E. Frank, M.A. Hall, Data Mining: Practical Machine Learning Tools and Techniques, Morgan Kaufmann, Burlington, MA, 2011.
[70] W. Yi, M. Lu, Z. Liu, Multi-valued attribute and multi-labeled data decision tree algorithm, Int. J. Machine Learn. Cybernet. 2 (2011) 67–74.
[71] Y. Zhao, Interactive Data Mining, VDM Verlag, 2008.
[72] Y. Zhao, Y. Yao, Interactive classification using a granule network, in: Proc. of the 4th IEEE Conf. on Cognitive Informatics, ICCI 2005, Irvine, CA, 2005, pp. 250–259.
[73] Y. Zhao, Y. Yao, On interactive data mining, in: J. Wang (Ed.), Encyclopedia of Data Warehousing and Mining, Idea Group, London, 2008, pp. 1085–1090.
[74] M. Žnidaršič, M. Bohanec, Automatic revision of qualitative multi-attribute decision models, Found. Comput. Decis. Sci. 32 (2007) 315–326.

Vedrana Vidulin is a researcher at the Department of Intelligent Systems at Jožef Stefan Institute, Ljubljana, Slovenia. She received her PhD degree at the Jožef Stefan International Postgraduate School, Ljubljana, Slovenia. Her research interests are related to data mining and machine learning, human–computer interaction, ambient intelligence and automatic web genre identification. Her major scientific achievement is the Human–Machine Data Mining method, which she has applied to several domains in the fields of macroeconomic research, automatic web genre identification and demography.

Marko Bohanec is a senior researcher at the Department of Knowledge Technologies, Jožef Stefan Institute, Ljubljana, Slovenia, and a professor of computer science at the University of Nova Gorica, Slovenia. His major research interests are related to decision support systems, data mining and, particularly, multi-criteria modeling, machine learning and the integration of data mining and decision support. He has developed a number of decision support tools and systems, such as DECMAK, DEX and DEXi, and applied them in management, economy, ecology, agronomy, medicine and healthcare.

Matjaž Gams is the head of Department of Intelligent Systems at Jožef Stefan Institute and a professor of computer science at the University of Ljubljana and Jožef Stefan International Postgraduate School (IPS), Slovenia. He received his degrees at the University of Ljubljana and IPS. He teaches or has taught on 10 faculties in Slovenia and Germany. His professional interests include intelligent systems, artificial intelligence, cognitive science, intelligent agents, business intelligence and information society. He is a member of the editorial boards of 11 journals and is the managing director of the journal *Informatica*. He is also a co-founder of various societies in Slovenia, such as the Engineering Academy, AI Society and Cognitive Society, and is the president or secretary of various societies, including ACM Slovenia. His major scientific achievement is the discovery of the principle of multiple knowledge.