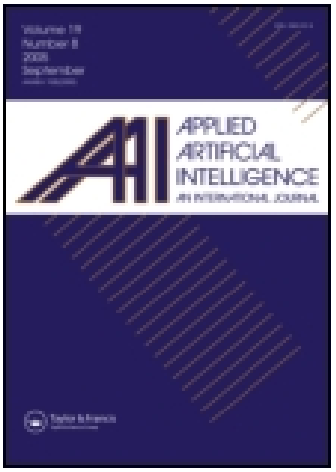

Applied Artificial Intelligence: An International Journal

Publication details, including instructions for authors and subscription information:
http://www.tandfonline.com/loi/uaai20

IMPACT OF HIGH-LEVEL KNOWLEDGE ON ECONOMIC WELFARE THROUGH INTERACTIVE DATA MINING

Vedrana Vidulin [a] & Matjaž Gams [a]

[a] Jožef Stefan Institute, Department of Intelligent Systems , Ljubljana, Slovenia

Published online: 18 Mar 2011.

To cite this article: Vedrana Vidulin & Matjaž Gams (2011) IMPACT OF HIGH-LEVEL KNOWLEDGE ON ECONOMIC WELFARE THROUGH INTERACTIVE DATA MINING, Applied Artificial Intelligence: An International Journal, 25:4, 267-291

To link to this article: http://dx.doi.org/10.1080/08839514.2011.559571

Applied Artificial Intelligence, 25:267–291, 2011

ISSN: 0883-9514 print/1087-6545 online
DOI: 10.1080/08839514.2011.559571

Taylor & Francis
Taylor & Francis Group

IMPACT OF HIGH-LEVEL KNOWLEDGE ON ECONOMIC WELFARE THROUGH INTERACTIVE DATA MINING

Vedrana Vidulin and Matjaž Gams
Jožef Stefan Institute, Department of Intelligent Systems, Ljubljana, Slovenia

□ *This paper describes a novel algorithm for finding the most important relations with the use of data mining. As an example application, the impact of high-level knowledge on economic welfare was analyzed. Our approach, based on interactive data mining, not only helps to discover the most relevant models, but also enables an evaluation of their relevance. The approach is specialized for the analysis of macroeconomic data that often contains incomplete and noisy attributes and, initially, complex relations, where several relations are statistically seemingly important, but only a few are indeed the most relevant. Although data mining algorithms are designed to detect the relevant attributes, irrelevant attributes often appear in the models due to chance-choice, reducing their quality and understandability, and consequently even leading to incorrect conclusions. We present our method and show its application at finding the most relevant relations between high-level knowledge and the state of development of a country.*

The primary task of data mining (DM) and machine learning (ML) is the automatic extraction of knowledge from data. Although some of the DM and ML methods output models in a human-understandable form (e.g., decision trees), these methods still lack an explanation as to why a specific model is the most relevant one. In other words, they lack an additional explanation that would convince us that we drew proper conclusions from the analysis. The field of interactive DM connects the fields of DM and human-computer interaction (HCI) to achieve the practical goal of developing methods to explain a computer's reasoning to a human and enabling the human to provide the feedback (Zhao and Yao 2008). The role of a computer is to perform complex computations and to explain the obtained results, while the human leads the DM process towards the correct and understandable domain models.

Address correspondence to Vedrana Vidulin, Jožef Stefan Institute, Department of Intelligent Systems, Jamova cesta 39, 1000 Ljubljana, Slovenia. E-mail: vedrana.vidulin@ijs.si

The initiative to connect the fields of ML and HCI was presented in a special issue of the journal Applied Artificial Intelligence "Machine Learning Meets Human-Computer Interaction" (1997). At that time, only one-third of the HCI researchers used ML methods (Moustakis and Herrmann 1997). In contrast, in the past few years DM and ML methods have been increasingly used in the HCI community for various applications, for example, for object recognition (Fails and Olsen 2003), e-mail categorization (Stumpf et al. 2009), affect detection (D'Mello and Graesser 2009) and sensory recommendation (Costello and McGinty 2009). Our paper is also concerned with the integration of ML and HCI, presenting an approach that is partially related to Stumpf et al. (2009). Similarly, the goal is to improve the relevance of a model constructed with DM methods through HCI, however, we use a different concrete method.

In this paper we focus on exploring the impact of high-level knowledge on the economic welfare of a country. Researchers in artificial intelligence (AI) have long been interested in this problem. One of the leading researchers in this area is Robert Trappl, who analyzed the impacts of AI-based technologies on society and the economy (Trappl 1986). Several of his optimistic predictions have already been realized. Now we live in the world where intelligent systems help us in everyday life (Jennings and Wooldridge 1995), from intelligent home appliances to product recommenders and tax advisors. They even help us with complex tasks such as determining the reputation of people we cooperate with electronically (Zacharia and Maes 2000). The internet itself can be seen as a major intelligent-services platform, offering new services such as e-learning (Baumgartner and Payr 1998) and significantly affecting the everyday life of citizens (Winston 1998). Our research is similar since we aim to determine the relations between knowledge and welfare. However, while the presented research of Trappl aims to determine the consequences of introducing AI-based technologies into society, we focus on the possible causes within the sectors that produce high-level knowledge and technology (including AI-based) with the highest impact on economic welfare, which is mainly represented as the growth of GNI per capita (Gylfason 2001) or GDP per capita (Keller 2006).

In general, the important influence of the R&D and higher education sectors on economic welfare was already acknowledged in the literature. The open question, however, remains, as to which segments of the two sectors have the highest impact on economic welfare. The segments deemed important are as follows: R&D expenditure, the number of researchers (Furman et al. 2002; Wang 2007), patents and academic publications (Furman et al. 2002; Varsakelis 2006; Wang 2007), share of GDP spent on higher education, percentage of R&D expenditure funded by private industry (Furman et al. 2002), achievements of students in mathematics

and science tests, the number of students in science higher education (Varsakelis 2006), public expenditure on education as a percentage of GNI, expected number of years of schooling for females (Gylfason 2001) and the enrollment rates in higher education (Keller 2006). A common feature of these approaches is that they are deductive in nature and only account for a subset of the segments in accordance with a specific theory. As a result, an important segment can be overlooked, simply because it is not a part of the theory. In contrast, we propose an inductive approach that accounts for all the available segments and extracts those that are the most important.

The power of interactive DM becomes obvious when mining incomplete and noisy data in complex domains, which is the case with macroeconomic data. Incomplete attributes represent the segments for which some of the countries did not report data. Noisy attributes in the complex domains represent those segments not directly related to the class. Standard DM methods are already designed to detect relevant attributes. However, irrelevant or somehow relevant attributes are sometimes included in the model, consequently reducing its quality and understandability. It has been shown that one random binary attribute can decrease the accuracy of a decision tree by 5 to 10% (John 1997). Incorrect models in turn lead to incorrect conclusions. To address the presented problems we propose an interactive DM method where the computer constructs the models, detects the irrelevant or less relevant attributes in the models, suggests corrections to the user and enables modifications of the model to achieve a better ratio between its quality and understandability.

Our approach belongs to the category of interactive pattern explanation and evaluation (Zhao and Yao 2008) where an interactive DM system constructs models and explains itself to the user, while the user makes the final judgment about the pattern's relevance. For the construction of models we selected two ML methods that produce easily understandable models in the form of decision and regression trees. The main advantage of trees lies in their comprehensibility, as demonstrated in similar domains, e.g., the analysis of historical data (Drummond et al. 2006) and international conflict data (Furnkranz et al. 1997; Druckman et al. 2005). The problems we addressed in more detail are how to integrate the power of DM and humans, i.e., how to explain the reasoning of a DM system to the user and how to collect the corrective feedback from the user.

Two types of explanation are described in the literature. Static explanations assume that the constructed model is a correct domain description. The goal is to explain which attributes participated in the model's decision to classify an instance into a specific category (Možina et al. 2004; Štrumbelj et al. 2009). In contrast, interactive explanations assume that the constructed model can be further improved. They are like static explanations with an added functionality to collect the corrective feedback from the user

(Stumpf et al. 2009; Kulesza et al. 2009). Existing interactive explanations are mainly oriented towards an improvement of the model's predictive performance. An explanation is based on an instance being classified, with the goal being to collect the corrections that will improve the model's predictive performance on future instances. The user does not need to be aware of the entire model, which can be complex and even contain less-important structures that only slightly improve the predictive performance. Our goal is different. Since our task is domain analysis, the constructed model should not only be of high quality, but also understandable to the user. Consequently, we present explanations oriented towards detecting and removing any spurious structure in the constructed model. Such corrections are directed towards improving both characteristics of the model.

Corrective feedback is usually collected in one of the following manners: first, by asking for more examples (Fails and Olsen 2003); second, by asking for more labels for unlabeled examples (Cohn et al. 1996); and third, by introducing domain knowledge into the DM process. The advantage of the last approach is that it directly affects the structure of the model. Therefore, the consequences of the corrections are immediately visible. Domain knowledge can be incorporated into the DM process by defining constraints, co-training ML-constructed and user-constructed models (Stumpf et al. 2009), or by the manual construction of models (Ware et al. 2001; Zhao and Yao 2005). The approach we propose is constraint based. In other words, we constrain the DM process by defining the parameters and attribute subsets of interest. In this manner, the user leads the DM towards understandable and high-quality domain models.

The rest of the paper is organized in five sections. First, we describe the interactive DM method that is specialized for the analysis of macroeconomic data. Second, we describe the data, followed by the analysis. The analysis reveals the relations between the high-level knowledge and the economic welfare, additionally showing the capability of our method to find high-quality and understandable models. The fourth section supplements the analysis by testing the ability of our method to find non-random patterns in data, additionally supporting our conclusions. Finally, the last section closes the paper with a discussion, ideas for future work and conclusions.

KNOWLEDGE ACQUISITION THROUGH HEURISTIC DATA MINING

The basic idea of our approach is to construct a large number of models with DM in a way that enables humans to extract the most relevant patterns – those that are understandable and of high quality. The task is

computationally very demanding because from n binary attributes it is possible to construct 2^{2^n} decision theories. The space of all the potential hypotheses for 100 binary attributes and a single binary class is therefore $2^{2^{100}}$. This number is far larger than the number of all the atoms in our universe, which according to Wikipedia is around 10^{90}, i.e., 2^{266}. Therefore, humans cannot successfully analyze hypotheses on their own.

Our approach is based on two assumptions. First, that from the enormous number of all hypotheses only a couple of them best represent the key relations in the domain. Second, that the search mechanism will discover relevant hypotheses and that humans will recognize the best of them. Indeed, this is our experience in recent years in most of the real-life domains describing economic and social relations.

The approach is based on an integration of human smartness and the brute force of computer DM methods. When DM methods perform a search, humans examine and evaluate the results, make conclusions and direct a new search. In this way, humans guide the DM towards relevant models, at the same time constructing an integrated conclusion from various solutions.

We use two basic heuristics. First, we examine the whole set of various parameters (typically algorithm parameters and attribute selection) to get a clue as to where the most interesting patterns might be. Second, as soon as an interesting pattern occurs, several heuristics are applied for cross-checking the relevance of the pattern. Relevant patterns are then stored and a new search begins until no major new relation is found for a while.

Although the approach is based on human judgment, commonly the most important patterns emerge quite evidently. In some cases there are several similar-quality patterns constructed with various attributes, meaning there is no dominant relation, while in other cases specific patterns significantly outperform the competing ones.

Our Knowledge Acquisition through Heuristic Data Mining (KAHDM) approach is based on the interaction model presented in Figure 1. In general, the interaction flows as follows. First, the human chooses an interesting subset of attributes, the DM method and the range of preferred parameters. Second, the system constructs a set of models in accordance with the user's preferences. Third, the user chooses one or a couple of candidates for the most relevant models, re-analyzing them by applying modifications and evaluating the results. The modification process is repeated until the user is not satisfied with the model, which is then stored and the conclusions integrated. Finally, the whole process is repeated until the user concludes that all the relevant patterns are extracted. Too many variations in the search options cause a combinatorial explosion; however,

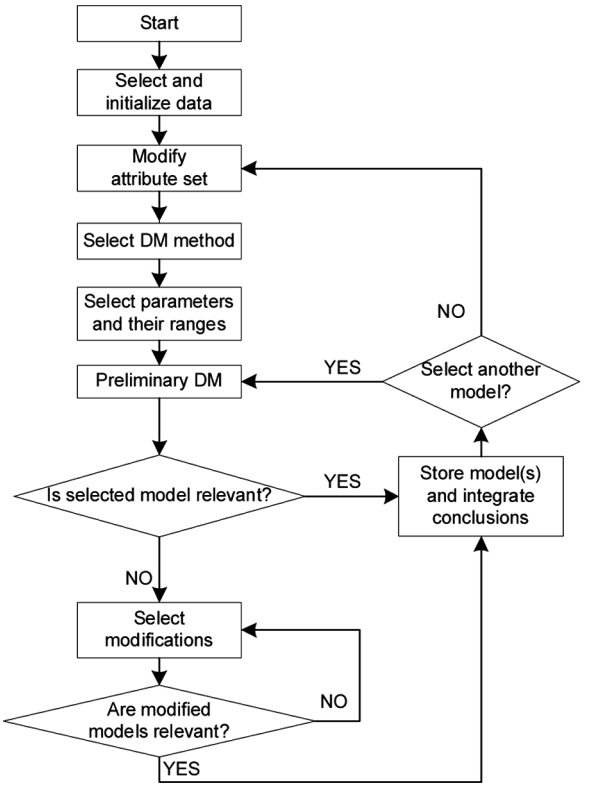

FIGURE 1 The interaction model of the KAHDM approach.

the search in our approach is guided by a human goal to verify an already-found pattern and supported with a specialized algorithm, implemented in an interactive computer system. As can be seen from examples in the rest of the paper, several variations can be quickly discarded as non-perspective. In the following paragraphs, the steps are described in more detail.

Select and initialize data. The user selects a data set.

Modify attribute set. This data preprocessing step is optional. The user can construct new attributes from the existing and/or select a subset of attributes.

Construct attributes. The attribute set is enriched with attributes derived from the existing attributes using operations like minimum, maximum and ratio. In this manner, relations become more explicit.

Select attributes. A subset of attributes can be selected manually based on expert knowledge, or automatically. Two automatic attribute selection tools were included in the system, which produce subsets representing the domain in two levels of detail. A small group of the most prominent

non-redundant attributes is obtained with the CFS subset selection method (Hall 1999). A larger set of attributes related to the class is obtained by measuring the information gain (IG) (Mitchell 1997) of the individual attributes and selecting those showing any relation with the class ($IG > 0$).

Select DM method. We included two DM methods that construct decision and regression trees. These models are directly comparable, facilitating the extraction of the most relevant patterns. For example, a pattern that appears in both the decision and regression tree, given the same data, considerably gains in relevance. For the induction of the trees we used algorithms from WEKA (Witten and Frank 2005): J48, the implementation of C4.5 (Quinlan 1993) for the induction of decision, and M5P (Quinlan 1992) for the induction of regression trees.

Select parameters and their ranges. The user selects the algorithm's parameters and the ranges in which the parameters are going to vary. In general, the understandability of the model is connected with its complexity, where less complex models are usually more understandable and vice versa. Therefore, we exploited those parameters that control the model's complexity. One parameter is the minimal number of instances per leaf (MNIL). For example, when set to 5, only those leaves that contain more than or equal to 5 instances are included in the tree. We varied this parameter between the default algorithm's value (4 for regression and 2 for decision trees) and 15. In the case of decision trees, the complexity was further controlled by setting the pruning procedure—standard C4.5 approach (STP) or the reduced-error pruning (REP). The other parameters are defined in the system manual.

Preliminary DM. The preliminary DM is the exploratory data-analysis phase, where the goal is to find a non-random pattern in the data (Mardia et al. 1979). A set of models is constructed with the selected DM method, one model for each parameter combination. The set is further refined by removing duplicate models. Then, the models are ranked in decreasing order of quality, and when several models are of equal quality, they are ranked in increasing order of complexity. In this manner, we facilitate access to the highest-quality models with the least complexity. We named this procedure **PRELIMINARY_DM**. The models are then presented to the user, who can browse them. When the user considers a certain model as relevant, he/she can store it. In contrast, when a model seems like a good candidate for the most relevant model, but the user doubts its relevance, he/she can re-examine it in further steps.

For measuring the model's quality we used ten-fold cross-validation (Kohavi 1995). Cross-validation is the most suitable solution when there is a limited amount of data, which is the case with macroeconomic data. In the case of decision trees the quality is expressed as the accuracy (ACC), and in the case of regression trees, as the correlation coefficient

(CC). The CC measures the correlation between the predicted and actual values. The complexity was expressed as the number of leaves, which corresponds to the number of relations within the tree.

Select modifications. The goal in this step is to (dis)confirm an already-found pattern (Mardia et al. 1979). At any time, the user can confirm the current model, but typically re-examination starts with a null hypothesis that the candidate model is irrelevant. The user can then select one or several modified models as a correction of the candidate model. The user can select between one of the five modifications: remove attributes, add attributes, construct concepts from attributes, remove concepts and add concepts.

Remove attributes. Typically, when a relevant attribute is removed and the model is reconstructed on the rest of the attributes, the quality of the model decreases. In contrast, when an irrelevant attribute is removed, the quality remains the same or even increases. A candidate model that contains only relevant attributes is relevant according to the quality criterion, i.e., the agreement with the data. It is up to the user to assess the model's understandability and to finally confirm or disconfirm its relevance. Furthermore, if the user decides to disconfirm the relevance of the candidate model, he/she can explore similar models in the vicinity of the candidate model and find a suitable correction.

The removed attributes are presented in the form of a graph (see Figure 4a). A node in the graph represents a removed attribute and the quality obtained after the attribute is removed and the model re-constructed using the rest of the attributes. By clicking on the node, the user can re-examine the re-constructed model. The hierarchical structure of the graph represents the removal of several attributes. The graph therefore represents the candidate model's relevance. It is up to the user to examine the graph and make the final judgments.

The graph can be constructed in an interactive, automatic or combined manner. In the interactive approach the user selects an attribute or a subset of attributes to be removed. The attributes are then removed and the nodes are added to the graph. The **REMOVE_ATTRIBUTES** procedure is presented in Figure 2. The procedure resembles the wrapper attribute selection approach (Kohavi and John 1997), but our goal is to present the relevance of the different attribute subsets to the user instead of finding the one best attribute subset. In the combined approach, a graph is first constructed automatically, followed by the interactive selection of attributes. In the examples we applied the combined approach to reassure the correct conclusions.

Add attributes. The idea behind adding attributes is to observe the role of a specific attribute or a subset of attributes in the process of pattern formation in the isolation of other attributes. For example, an attribute can

REMOVE_ATTRIBUTES *(data set, DM method, model, parameters and their ranges)*
  for *each attribute* a_i *in the model*
    Remove attribute a_i
    Perform **PRELIMINARY_DM** *(data set -* a_i*, DM method, parameters and their ranges), returning just the top-ranked model*
    if *quality of the top-ranked model is worse than the previous one*
    then
      Construct node n_i *from the top-ranked model, its quality and the deleted attribute* a_i
      Add n_i *to the graph*
      REMOVE_ATTRIBUTES *(data set -* a_i*, DM method, top ranked model, parameters and their ranges)*
  end for
  return *graph*

FIGURE 2 The algorithm describing the **REMOVE_ATTRIBUTES** procedure.

appear in a cross-validation fold, consequently influencing the quality of the model. At the same time, this attribute might not appear within the constructed model. Here, we are primarily interested in the quality of those attributes that directly modify the structure and therefore influence the final conclusions. When the subset of the attributes from the candidate model in isolation produces an equal quality model, and the model is understandable to the user, then the relevance of the candidate model is confirmed. In contrast, the user can select one or several corrections to the candidate model, without the irrelevant attributes included.

The attributes are added in a similar manner as they are removed. The main difference is that automatic graph construction is made with the **ADD_ATTRIBUTES** procedure. It constructs a graph (see Figure 4b) in the reverse order to the **REMOVE_ATTRIBUTES** procedure, starting from a predefined set of attributes and adding new attributes while the quality increases. If the user does not define the initial attribute set, then the procedure starts from the empty set.

Construct concepts. Certain attributes represent the same semantic category, however, expressed in a different manner. For example, "GERD per capita (PPP$)" and "GERD as % of GDP" both represent the level of investment in R&D, but expressed in different quantities. Such attributes are correlated and when one is removed the other usually takes its place. Therefore, we provided a tool for grouping such attributes into concepts. The concepts are denoted with small capital letters. Each concept is followed by the number of attributes contained within the concept.

Remove concepts. The procedure is basically the same as the **REMOVE_ATTRIBUTES** procedure, with the difference being that the subsets of the attributes marked as concepts are removed together and represented with a single node.

Add concepts. Similar to remove concepts, only this time the concepts are added.

DATA

We collected data representing the high-level knowledge sectors from several statistical databases provided by the following: UNESCO Institute for Statistics,[1] USAID – Global Education Database,[2] and WIPO[3]. In total, 108 numerical attributes were collected: 48 describing inputs (personnel and financial resources) and outputs of the R&D sector, and 60 describing inputs to the higher education sector. The data was gathered for 167 countries, thus the learning matrix consists of 108 columns and 167 rows. The data and the description of attributes can be obtained from: http://dis.ijs.si/Vedrana/economic-analysis.htm.

The economic welfare is represented by the "GNI per capita" attribute, calculated according to The World Bank Atlas method. GNI stands for the *Gross National Income* and represents the total value of goods and services produced within a country (Black et al. 2009). We collected the "GNI per capita" from The World Bank database[4] in both numerical and discrete form. The numerical form is expressed in US$, while the discrete form represents the official classification of the countries into income levels: low—$745 or less, middle—$746–9,205, and high—$9,206 or more. From the total of 167 countries, 50 belong to the low, 79 to the middle and 38 to the high income groups.

IMPACTS OF HIGH-LEVEL KNOWLEDGE

In this section, we present the analysis of the macroeconomic data with the KAHDM method. It is separately applied in two sessions: first on 48 R&D attributes and second on 60 higher education attributes. It should be noted that we present only the relevant findings, while the majority of the tests with less significant results are omitted due to a lack of space.

Impact of the R&D Sector

Regression Trees

Preliminary DM. The preliminary DM resulted in 4 trees, from which we selected the tree in Figure 3, constructed with the parameter MNIL 5. The tree indicates that just the level of investment in R&D represents the key factor in differentiating countries. GERD stands for the *Gross Domestic Expenditure on R&D*, denoting the expenditure on R&D performed on

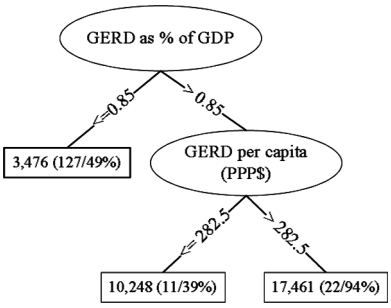

FIGURE 3 The regression tree constructed on 48 R&D attributes; CC 0.73.

the national territory during a year (Black et al. 2009). Those countries that invest in R&D with less than or equal to 0.85% of their GDP (leftmost leaf) have an average GNI per capita of 3,476 US$. A total of 127 countries (the first number in brackets) or 79% of them conform to this description. The deviation around the average value is 49% (the second number in brackets—representing the root-mean-squared error divided by the global absolute deviation). In comparison to The World Bank's categories, this group includes the low- (745 US$ or less) and the middle-income (746–9,205 US$) countries, meaning that the high-income countries (9,206 US$ or more) invest in R&D more than 0.85% of their GDP (right subtree). High-income countries are even further differentiated according to the level of investment in R&D (stated in PPP$) into two groups, again showing that a higher level of investment in R&D leads to an even higher income. *PPP*$ stands for purchasing power parity in American dollars. The statement of GERD per capita in PPP$ allows for fair comparisons between different countries. The obtained tree is understandable and of reasonable quality. However, there is an open question: is GERD the only important segment?

Remove attributes. The graph of removed attributes is presented in Figure 4a. Note that the actually constructed graph is quite a bit larger; however, only the parts relevant to the discussion are shown. "GERD per

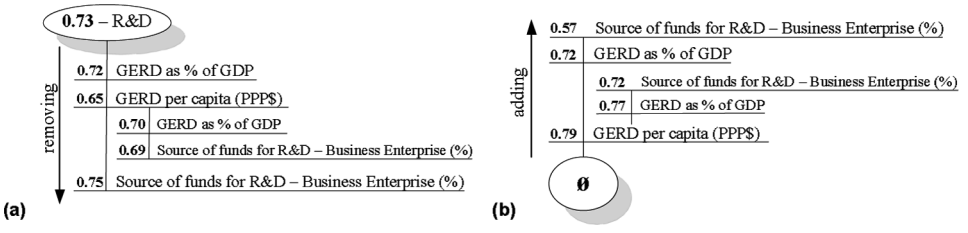

FIGURE 4 The relevance of R&D attributes – regression trees.

capita" appears as the best single attribute, since the highest fall in quality is observed when it is removed (from 0.73 to 0.65). The removal of other attributes in combination with it did not result in any further fall in the quality, indicating it is indeed an important attribute. Removing both GERDs results in a decrease from 0.73 to 0.70, which seems a little confusing, since they both appear in Figure 3 and the decrease is smaller than removing only the "GERD per capita". Another interesting finding is the appearance of the tree with a CC of 0.75 after the removal of the "Source of funds for R&D—Business Enterprise (%)" attribute. Its quality is higher than the quality of the candidate tree (0.73). The analysis showed that both trees represent the same pattern; however, in the cross-validation for the candidate tree, out of 10 trees 3 contained the source of funds attribute, consequently lowering the CC. Since the quality increased when the source of funds attribute was removed, this indicates that DM overestimates its importance.

Add attributes. In Figure 4b "GERD per capita" again appeared as the most relevant attribute—the highest-quality tree (0.79) was constructed just on this attribute, while additional attributes resulted in degradation. The highest-quality tree is presented in Figure 5, showing a similar relation to the tree in Figure 3—a higher level of investment determines a higher income. Since "GERD per capita" is measured in PPP$, it corresponds to the percentage and not to absolute numbers.

Conclusions. A higher level of investment in R&D leads to better economic welfare, supported by a CC of 0.79. Since "GERD per capita" and "GERD as % of GDP" represent the same semantic category, only expressed in different quantities, the trees in Figure 3 and Figure 5 support the same conclusion. Other attributes seem less relevant.

Decision Trees

Preliminary DM. Preliminary DM resulted in 20 trees, from which we selected the tree in Figure 6 constructed with the parameters MNIL 3

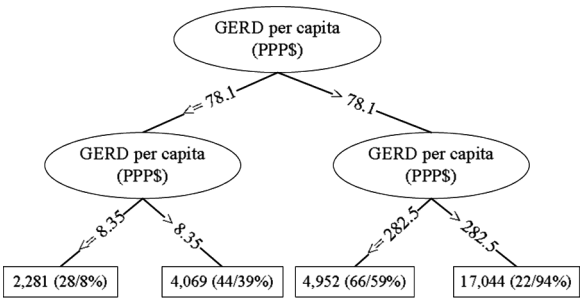

FIGURE 5 The regression tree constructed on "GERD per capita" attribute; CC 0.79.

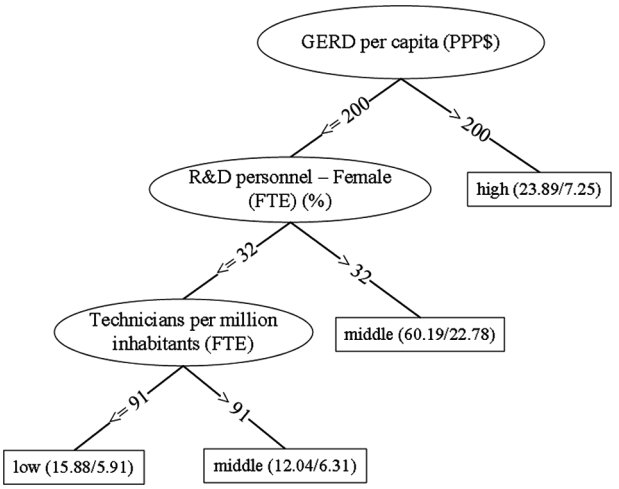

FIGURE 6 The decision tree constructed on 48 R&D attributes; ACC 63%.

and REP. The topmost relation in the tree is similar to those in the regression trees—countries with a higher level of investment in R&D also have a higher income. In the right-up leaf of the tree, "high" denotes the countries that invest more than 200 PPP$ in R&D. The first number in brackets (23.89) represents the number of countries that reached the leaf and the second number (7.25) those of the 23.89 countries that belong to classes other than the majority class. The number of countries is expressed in decimals to account for the missing values (Quinlan 1993). A further differentiation between the countries was made based on the position of females in the R&D sector, which seems a bit suspicious; therefore, further analyses are needed.

Construct concepts. From the experiments with attributes we noticed that the connected attributes substitute for each other. For example, when "GERD per capita" was removed, "GERD as % of GDP" took its role as a root of the tree. To facilitate the analysis we grouped the connected attributes into the concepts. In total, 18 concepts were created.

Remove and add concepts. The graph in Figure 7 was obtained by fusing two graphs constructed by removing and by adding concepts to stress only the important findings. The level of investment in R&D again turned out to be highly important, as seen in two ways. First, when the GERD concept was removed, ACC fell by 3 percentage points (PP). In contrast, when other concepts were individually removed, none of them caused a fall in ACC, including R&D PERSONNEL – FEMALE, indicating that the initial DM overestimated this relation. Second, the tree constructed just on the GERD concept (last line in Figure 7) has a 3 PP higher ACC than the

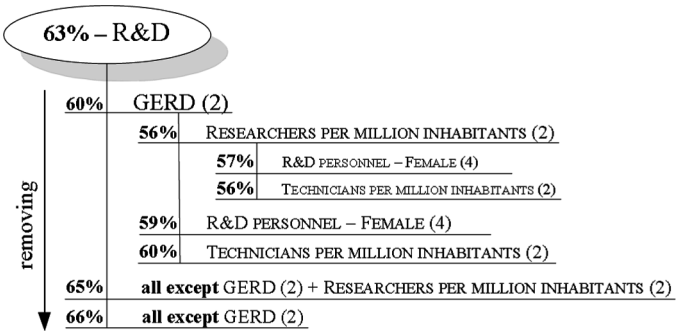

FIGURE 7 The relevance of R&D concepts – decision trees.

highest-quality tree constructed on all 48 R&D attributes. All the other concepts seem less relevant. The R&D PERSONNEL – FEMALE introduces only a 1 PP decrease in ACC if removed after the GERD, further confirming the non-top relevance. For example, when RESEARCHERS PER MILLION INHABITANTS (represented with two attributes in FTE (full-time equivalent) and HC (head count)) was removed, the ACC fell by another 4 PP. However, the tree constructed on the GERD concept is of higher quality than the tree constructed on the combination of GERD and RESEARCHERS PER MILLION INHABITANTS concepts. The analysis of other combinations did not result in a combination that would outperform the GERD concept.

The highest-quality tree constructed on the GERD concept with the parameters MNIL 3 and REP is presented in Figure 8. One can see a slightly surprising relation in the deepest part of the tree: the low-income countries have a "GERD as % of GDP" higher than 0.15, while there are middle-income countries that invest less of their GDP into R&D. At first glance, this disagrees with the conclusion that a higher level of investment in R&D is connected to better economic welfare. A further analysis showed that the subtree is not the result of an error due to, e.g., missing values. These countries are rich in natural resources and therefore "jump" on the ladder. The DM method correctly pointed out this exception and analyses further confirmed the relevance of the overall relation. Therefore, we selected the tree as a correction of the candidate tree.

Conclusions. Both regression and decision trees show that the level of investment in R&D represents the key factor in differentiating countries according to their economic welfare—GERD on its own is the most relevant indicator of a country's welfare. Other segments have some influence on welfare, which is less significant and should be treated with caution. Their relevance has yet to be proven beyond doubt.

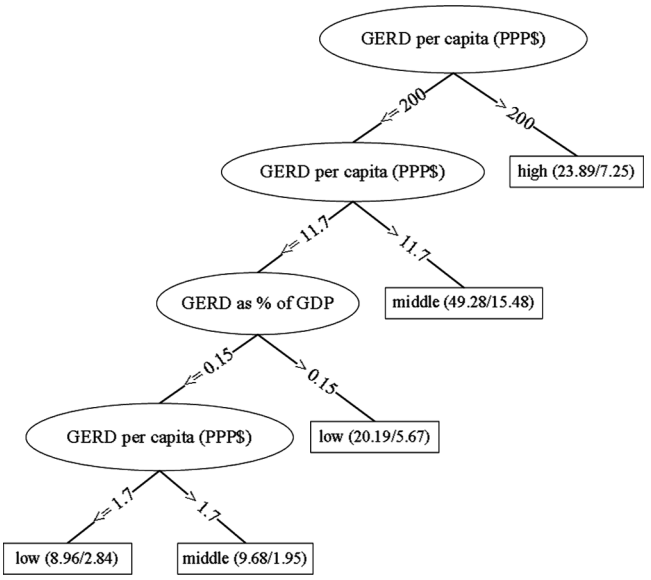

FIGURE 8 The decision tree constructed on the GERD concept; ACC 66%.

Decision Trees from the Modified Attribute Set

Construct attributes. A total of 33 new attributes were constructed based on the observations. For example, an attribute "Sector investing the most in R&D" was constructed by finding the maximum between six "Source of funds for R&D" attributes. Accordingly, an attribute takes one of six values: business enterprise, government, higher education, private non-profit, abroad and N/A (not available or not known distribution).

Select attributes. The attribute construction step was followed by the attribute selection step to reduce the number of irrelevant attributes. The first stage was the selection of attributes based on expert knowledge. In total, 31 attributes were removed. On the rest of 50 attributes we applied two attribute selection methods. The highest-quality tree constructed on attributes selected with the CFS subset selection method had an ACC of 65%. In contrast, the highest-quality tree constructed on attributes selected with the IG selection method had an ACC of 69%. We decided to exploit the second attribute selection method since it results in higher-quality trees.

Preliminary DM. The preliminary DM resulted in 16 trees, from which we selected the tree in Figure 9 constructed with the parameters MNIL 5 and STP. The tree confirms that a high level of investment in R&D is important for better economic welfare. High-income countries are characterized as those who invest more than 105.5 PPP$ per capita in R&D, while low-income countries as those who invest less than or equal to 10.8 PPP$. Common sense tells us that the right subtree is misleading. The values of

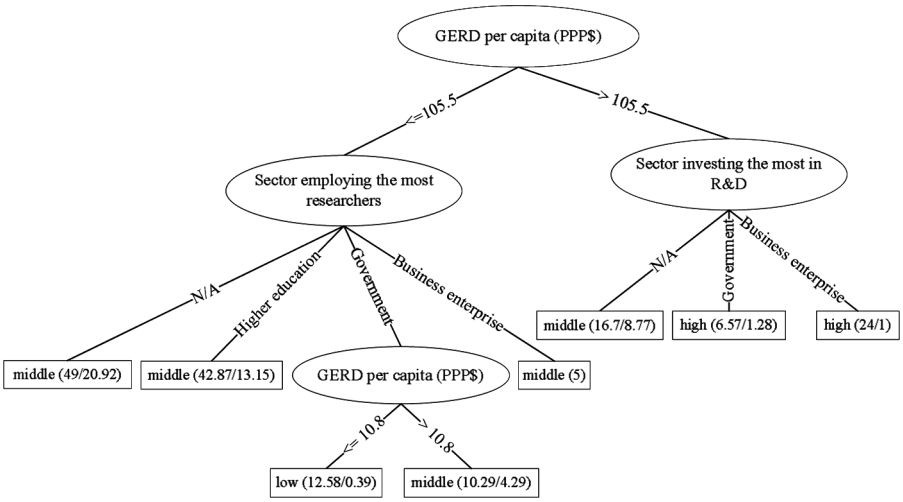

FIGURE 9 The decision tree constructed on the modified set of R&D attributes; ACC 67%.

"GERD per capita" and "Sector investing the most in R&D" (N/A branch) are missing for some countries, and the right subtree would be better represented by a single leaf "high".

Construct concepts. The scheme used to group the modified attribute set into concepts was similar to the scheme used to group the original attribute set, with the difference that all the attributes representing R&D personnel were grouped into a single concept. These attributes did not emerge as relevant in the previous analysis steps leading to such a decision. In total, 7 concepts were constructed.

Remove and add concepts. The graphs of removed and added concepts are presented in Figure 10. Because GERD had already proved important it was added in the initial concept set. Besides the level of investment in R&D, i.e., GERD (fall in ACC of 6 PP), two concepts appeared relevant: SECTORS EMPLOYING R&D PERSONNEL (4 PP fall and 2 PP gain) and R&D PERSONNEL (1 PP fall and 1 PP gain). To clarify which segments are indeed relevant we returned to the attribute level analysis. A further analysis is concentrated only on the three relevant attributes constituting the potentially relevant concepts. Note that we, in a way, redo the analysis of attributes (small letters), but this time with the original and additional attributes, because the current analysis pointed out some additional candidate relations.

Add attributes. The graph in Figure 11 shows that the most relevant are the GERD attributes in combination with the "Sector employing the most R&D personnel" attribute (ACC 71%). The attributes within the R&D PERSONNEL concept did not appear as highly relevant—each time when added they reduced the ACC.

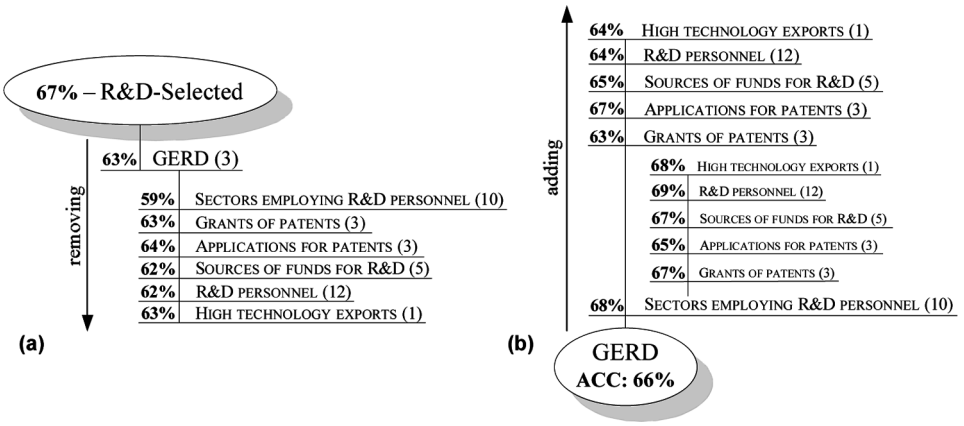

FIGURE 10 The relevance of concepts constructed from the modified set of R&D attributes.

The highest-quality tree constructed with the parameters MNIL 4 and STP is presented in Figure 12. The modifications did improve the quality, but questions persist. First, if we eliminate the N/A branch, the right subtree would be substituted with the leaf "high"; second, the same would happen with the leftmost subtree, which would be substituted with the leaf "low". Further analyses confirmed this suspicion.

Conclusions. The analysis again confirmed the level of investment in R&D (GERD) as the most relevant indicator of economic welfare; the only one in its league. But this time, an observation of the distribution of the left and right subtrees in Figure 12 indicated that low-income countries have most of the researchers in the government, while in the developed countries most of them are in business enterprises.

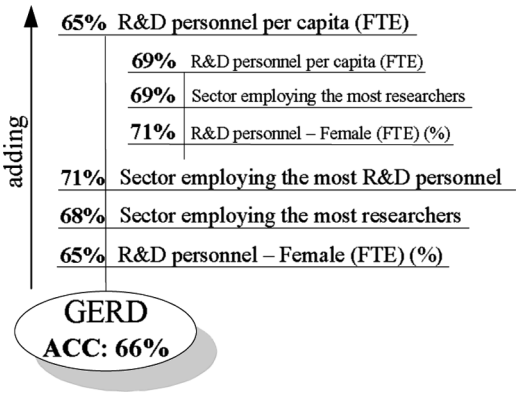

FIGURE 11 The relevance of attributes from the modified set of R&D attributes.

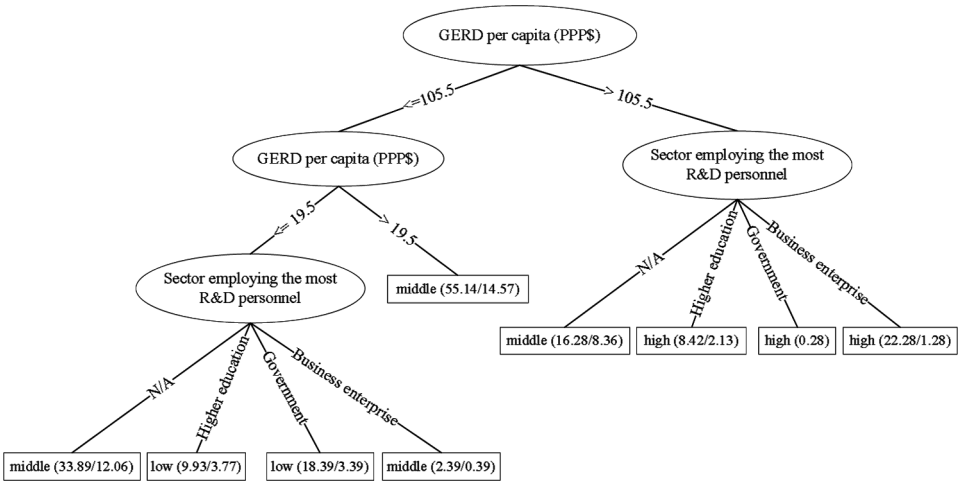

FIGURE 12 The decision tree constructed on the GERD and "Sector employing the most R&D personnel" attributes; ACC 71%.

Discussion

Several R&D segments (i.e., attributes and concepts) that influence economic welfare were detected through the analysis of the constructed trees. We divide them into three relevance levels: the first level contains those segments that consistently appeared throughout the trees and are valid for the majority of countries; the second level contains those segments that often modified the quality of the trees, while they only occasionally modified the tree's structure; the third level contains all the others.

Only the level of investment in R&D (GERD) belongs to the first category. Its importance was acknowledged in the economic literature. It is generally used as a control variable (Varsakelis 2006) to examine how successful the proposed method is at detecting relevant segments.

There are only three segments that appear in the second category, as can be observed, e.g., from Figure 10: patents, high-technology exports and the sector employing the most R&D personnel. The literature also supports these conclusions, e.g., (Furman et al. 2002).

Some segments belonging to the third level appeared in the analysis, e.g., the percentage of women researchers. Analyses sometimes consider them as relevant; however, verifications revealed that their influence is weak.

In the end, it is worth noting that it is difficult to directly compare our results with the results of other studies since they used different classes. While we used GNI per capita as the class, other studies accounted for the number of patents or the number of academic publications, thus

searching for which direction to follow to obtain more publications or patents. This can explain why some of the conclusions are different.

Impact of the Higher Education Sector

The higher education sector data was analyzed in a similar manner as the R&D sector data. However, preliminary DM with regression trees produced complex trees with the highest observed CC of 0.52 and a further analysis did not result in considerable improvements. Second, with decision trees we constructed 11 new attributes and tested several attribute selection methods, but the highest observed ACC was 74%, which is 2 PP worse than the highest observed ACC obtained with the original attributes. Therefore, the first impression was that the relations in this domain are of a different nature. Furthermore, analyses of the higher education sector are presented in less detail than in the R&D sector.

Decision Trees

Preliminary DM. Preliminary DM resulted in 14 trees from which we selected the one in Figure 13 constructed with the parameters MNIL 11 and STP. The tree contains two understandable and two problematic relations. The first understandable relation states that a high level of enrolment leads to better economic welfare. Enrolment is represented by the "*Gross enrolment ratio*—ISCED 5 and 6" (GER-Total) attribute. This attribute

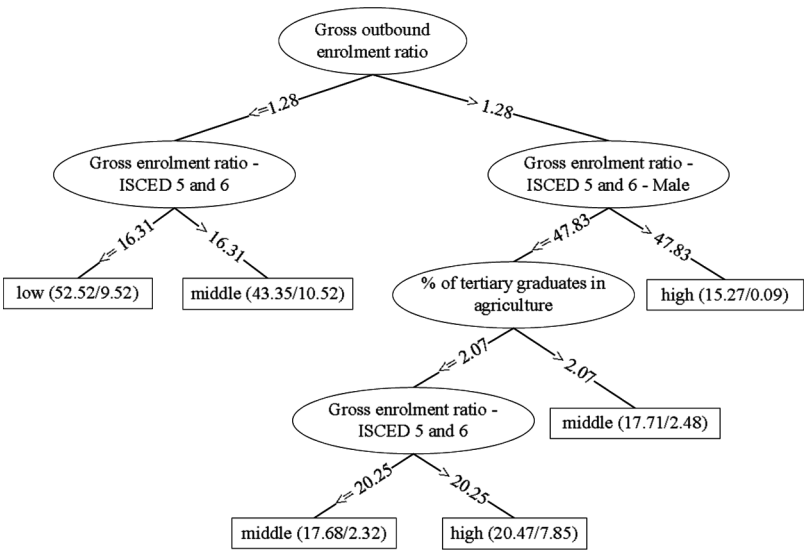

FIGURE 13 The decision tree constructed on 60 higher education attributes; ACC 74%.

accounts for students enrolled in higher education, regardless of their age, expressed as a percentage of the population in the five-year age group following on from leaving secondary school. *ISCED* 5 denotes the first stage of higher education, while ISCED 6 represents programs leading to the award of an advanced research qualification (UNESCO 2006). In other words: the more most educated citizens there are the better. The second understandable relation states that a higher mobility of students leads to better economic welfare. Student mobility is represented by the "*Gross outbound enrolment ratio*" (GOER) attribute, denoting students from a country studying abroad as a percentage of the population of higher education student age in that country. The relation straightforwardly differentiates between high- and low-income countries, while middle-income countries are somewhere in between, showing both lower and higher student mobility. We can state that middle-income countries are in a transition phase from lower to higher student mobility.

There are two suspicious attributes within the tree for which the relation remained undefined. First, the "Gross enrolment ratio—ISCED 5 and 6—Male" (GER-Male) subtree divides approximately the same number of high-income countries between both branches. The same happened with the "% of tertiary graduates in agriculture" subtree, only this time with the middle-income countries. A further analysis was needed to check these relations.

Construct concepts. Experiments with attributes showed that three GER attributes (GER-Total, GER-Male and GER-Female) substitute each other. Therefore, we grouped them into the GROSS ENROLMENT RATIO (GER) concept. Other similar groupings were not observed.

Remove and add concepts. In Figure 14, the best combination of attributes is the combination of the GER and GOER attributes (4 PP fall and 2 PP increase in ACC in comparison to the candidate tree), which results in a tree with an ACC of 76%. The graphs did not support the "% of tertiary graduates in agriculture" attribute. The removal and addition of other attributes did not result in further significant improvements.

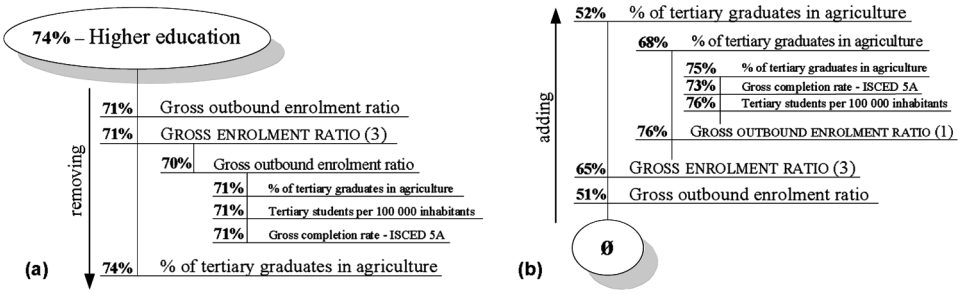

FIGURE 14 The relevance of higher education concepts and attributes.

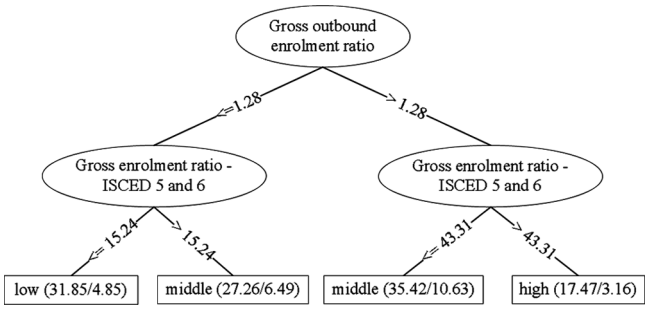

FIGURE 15 The decision tree constructed on the GER concept and the GOER attribute; ACC 76%.

The highest-quality tree constructed with the parameters MNIL 12 and REP is presented in Figure 15. The tree further clarifies conclusions based on the analysis of the candidate tree; this time representing only the important relations. The tree made clear distinctions between the levels of enrolment in countries with different incomes: low $\leq$15%, middle 15–43% and high >43%.

Conclusions. Results of the analysis indicate two important relations: higher participation in higher education and better student mobility leads to better economic welfare.

Discussion

The analysis showed that two segments of the higher education sector have primary importance for economic welfare. First, a higher participation in higher education leads to better economic welfare. Having an educated population positively influences not only the creation of new knowledge and technologies, but also the better exploitation of the new technologies. The importance of this segment was also recognized by Keller (2006), again confirming the ability of our method to find relevant relations. Second, with the help of our method we discovered that a higher mobility of students is also very important for better economic welfare. This relation was not directly discussed in the presented related work. In the analysis we considered two types of mobility: studying in foreign countries and students from foreign countries that study in the country of interest. The results showed that it is important to stimulate students to spend certain amount of time studying abroad, transferring new knowledge back to their home countries. The two relations belong to the first level of relevance.

Some other segments like "% of tertiary graduates in agriculture" emerged during preliminary DM, but additional analyses showed that they are level-three segments according to our categorization and are not discussed further.

We were surprised that none of the attributes representing the level of investment in higher education appeared important, especially since those attributes were denoted as relevant by other studies. Indeed, it seems reasonable that a greater percentage of highly educated people can only be achieved with a higher investment, but our analyses reveal an important difference between the two. This issue should be analyzed in more detail.

EVALUATION

Relevant trees should not only be understandable and of high quality, but should also represent non-random patterns in data to support our conclusions. To this end, we evaluated the trees stored during the analysis by comparing them to a random tree. A tree is considered as non-random when it is significantly better than a random tree. In the case of decision trees, we consider as random the tree that always returns the majority class (in our case "middle") and in the case of regression trees, if it returns the mean of the actual values (in our case 6258).

Table 1 presents the comparison of relevant decision trees with the baseline. The increase in quality varies from 19 to 29 PP, which is significantly different from the baseline, strongly indicating that the effort in using our approach paid off. Table 2 presents the comparison of the relevant regression trees with the baseline. We did not straightforwardly compute the differences in CC since they are of different signs. Considering that "negative values should not occur for reasonable prediction models" (Witten and Frank 2005), it is clear that our method was able to find non-random patterns in the data.

When the difficulty of the domain is taken into account, the findings are supported with regression and decision trees of considerable quality.

TABLE 1 Evaluation of Decision Trees

Decision trees	Baseline	ACC	Diff. (PP)
GERD (Figure 8)	47%	66%	19
GERD & Sector (Figure 12)		71%	24
GER & GOER (Figure 15)		76%	29

TABLE 2 Evaluation of Regression Trees

Regression trees	Baseline	CC
R&D (Figure 3)	−0.13	0.73
GERD per capita (Figure 5)		0.79

The main source of difficulty is the considerable amount of missing values scattered all over the data and the complexity of the task.

CONCLUSIONS AND DISCUSSION

From the economic perspective, several interesting relations were detected. For better economic welfare, it is very important to stimulate students to enroll in higher education. Furthermore, better programs for student mobility should be developed to stimulate the exchange of high-level knowledge. This seems understandable, since having a well-educated population does not only positively affect the production of high-level knowledge, but also the consumption of the products and services of high-level knowledge. In contrast, there is only one important step in R&D for a country to progress: raise the level of investment. However, several other segments are also important at level two, such as promoting patenting and lifting governmental control of science.

From the methodological perspective, we have presented a new method—Knowledge Acquisition through Heuristic Data Mining (KAHDM). The essential advantage is based on the interaction between the two most advanced information machines: the brute force of computers (enriched with AI DM) and human insight and comprehension. The implemented interactive system constructs a set of decision and regression trees and explains its reasoning to a human. A human leads DM with the goal being to find high-quality and understandable relations in the macroeconomic data. In comparison to the standard approach used in economic analyses, our approach offers several advantages. First, the DM is data driven, not theory driven, as in the majority of the existing methods. But our method also makes it possible to combine human theories with the data, providing a tool to thoroughly test and verify potential relations. Second, during the model-construction process it analyzes the impact of a large number of attributes at once. Finally, its results are easily comprehensible, even to non-experts.

There are two issues to be considered in our approach. First, is the method stable or will the results vary depending on a particular human performing the KAHDM? Regarding the first issue, the subjectivity can be eliminated by publishing the data and results and enabling access through the internet (see http://dis.ijs.si/Vedrana/economic-analysis.htm).

The second issue is which type of relation was indeed observed—X implies Y or Y implies X? Does more investment in R&D actually cause countries to progress faster or is it just a side-effect of the developed countries that they spend more on R&D? To be fair, the trees and analyses in this paper do not indicate the type of the relation. However, it is highly unlikely that such a strong relation would not be mutual, acting in both directions.

To evaluate these relations in quantitative ways, other methodologies are more appropriate than KAHDM.

Finally, it depends on human ingenuity to accept or reject any conclusion, however supported by statistics or any other formal method. By observing not only one decision tree in one DM set-up, but thousands of them and giving an interactive tool to verify the hypotheses enabling the human mind to integrate conclusions from thousands of constructed trees, the published relations emerged as relevant in a kind of real-life way, rather than in a formal way.

NOTES

1. http://www.uis.unesco.org
2. http://ged.eads.usaidallnet.gov
3. http://www.wipo.int
4. http://www.worldbank.org

REFERENCES

Baumgartner, P., and S. Payr. 1998. Educating the Knowledge Worker in the Information Society. In *Teleteaching'98—Proceedings of the 15th IFIP World Computer Congress*, ed. G. B. Davis. Vienna: Austrian Computer Society. 109–118.

Black, J., N. Hashimzade, and G. Myles. 2009. *A dictionary of economics*. New York: Oxford University Press.

Cohn, D. A., Z. Ghahramani, and M. I. Jordan. 1996. Active learning with statistical models. *Journal of Artificial Intelligence Research* 4:129–145.

Costello, E., and L. McGinty. 2009. Supporting User Interaction in Flavor Sampling Trials. In *Proceedings of Workshop on Intelligence and Interaction, IJCAI 2009*, eds. J. Fogarty, E. Horvitz, A. Kapoor, and D. Tan. Pasadena, CA.

D'Mello, S., and A. Graesser. 2009. Automatic detection of learner's affect from gross body language. *Applied Artificial Intelligence* 23 (2): 123–150.

Druckman, D., R. Harris, and J. Fürnkranz. 2006. Modeling international negotiation: Statistical and machine learning applications. In *Programming for peace: Computer-aided methods for international conflict resolution and prevention*, ed. R. Trappl. The Netherlands: Springer. 227–250.

Drummond, C., S. Matwin, and C. Gaffield. 2006. Inferring and revising theories with confidence: Analyzing bilingualism in the 1901 Canadian Census. *Applied Artificial Intelligence* 20:1–33.

Fails, J. A., and D. R. Olsen. 2003. Interactive Machine Learning. In *Proceedings of the 8th International Conference on Intelligent User Interfaces*, eds. D. Leake, L. Johnson, and E. Andre. Miami, FL. New York: ACM. 39–45.

Furman, J. L., M. E. Porter, and S. Stern. 2002. The determinants of national innovative capacity. *Research Policy* 31:899–933.

Furnkranz, J., J. Petrak, and R. Trappl. 1997. Knowledge discovery in international conflict databases. *Applied Artificial Intelligence* 11:91–118.

Gylfason, T. 2001. Natural resources, education, and economic development. *European Economic Review* 45:847–859.

Hall, M. A. 1999. Correlation-based Feature Selection for Machine Learning. PhD Thesis. Hamilton. New Zealand.

Herrmann, J., and V. S. Moustakis. 1997. Applied artificial intelligence: Machine learning meets human-computer interaction. *Applied Artificial Intelligence* 11(7&8).

Jennings, N. R., and M. Wooldridge. 1995. Applying agent technology. *Applied Artificial Intelligence* 9: 357–369.

John, G. H. 1997. Enhancements to the Data Mining Process. Ph.D. Thesis. Computer Science Department, Stanford University, Stanford.

Keller, K. R. 2006. Investment in primary, secondary, and higher education and the effects on economic growth. *Contemporary Economic Policy* 24 (1): 8–34.

Kohavi, R. 1995. A Study of Cross-Validation and Bootstrap for Accuracy Estimation and Model Selection. In *Proceedings of IJCAI*. Montreal, Quebec, Canada. 1137–1145.

Kohavi, R., and G. H. John. 1997. Wrappers for feature subset selection. *Artificial Intelligence* 97:273–324.

Kulesza, T., W.-K. Wong, S. Stumpf, S. Perona, R. White, M. M. Burnett, I. Oberst, and A. J. Ko. 2009. Fixing the Program My Computer Learned: Barriers for End Users, Challenges for Machine. In *Proceedings of Conference on Intelligent User Interfaces (IUI)*, eds. C. Conati, M. Bauer, N. Oliver, D. Weld. Sanibel Island, FL. New York: ACM.

Mardia, K. V., J. T. Kent, and J. M. Bibby. 1979. *Multivariate analysis*. London: Academic Press.

Mitchell, T. M. 1997. *Machine learning*. McGraw-Hill.

Moustakis, V. S., and J. Herrmann. 1997. Where do machine learning and human-computer interaction meet? *Applied Artificial Intelligence* 11 (7&8): 595–609.

Možina, M., J. Demšar, M. Kattan, and B. Zupan. 2004. Nomograms for Visualization of Naive Bayesian Classifier. In *Proceedings of 8th European Conference on Principles and Practice of Knowledge Discovery in Databases*, eds. J.-F. Boulicaut, F. Esposito, F. Glannotti, D. Pedreschi. Pisa, IT. Berlin: Springer. 337–348.

Quinlan, J. R. 1992. Learning with Continuous Classes. In *Proceedings of 5th Australian Joint Conference on Artificial Intelligence*, eds. A. Adams and L. Sterling, Hobart, Tasmania. Singapore: World Scientific. 343–348.

Quinlan, J. R. 1993. *C4.5: Programs for machine learning*. San Mateo, CA, USA: Morgan Kaufmann.

Stumpf, S., V. Rajaram, L. Li, W. Wong, M. Burnett, T. Dietterich, E. Sullivan, and J. Herlocker. 2009. Interacting meaningfully with machine learning systems: Three experiments. *International Journal of Human-Computer Studies* 67 (8): 639–662.

Štrumbelj, E., I. Kononenko, and M. Robnik Šikonja. 2009. Explaining instance classifications with interactions of subsets of feature values. *Data & Knowledge Engineering* 68 (10): 886–904.

Trappl, R. 1986. *Impacts of artificial intelligence*. Elsevier Science.

UNESCO. 2006. *ISCED 1997—International Standard Classification of Education*.

Varsakelis, N. C. 2006. Education, political institutions and innovative activity: A cross-country empirical investigation. *Research Policy* 35:1083–1090.

Wang, E. C. 2007. R&D Efficiency and economic performance: A cross-country analysis using the stochastic frontier approach. *Journal of Policy Modeling* 29 (2): 345–360.

Ware, M., E. Frank, G. Holmes, M. Hall, and I. H. Witten. 2001. Interactive machine learning: Letting users build classifiers. *International Journal of Human-Computer Studies* 55:281–292.

Winston, B. 1998. *Media technology and society: A history: From the telegraph to the Internet*. London, UK: Routledge.

Witten, I. H., and E. Frank. 2005. *Data mining—practical machine learning tools and techniques*. San Francisco, CA, USA: Elsevier Science.

Zacharia, G., and P. Maes. 2000. Trust management through reputation mechanisms. *Applied Artificial Intelligence* 14:881–907.

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

Zhao, Y., and Y. Yao. 2008. On interactive data mining. In *Encyclopedia of data warehousing and mining*, ed. J. Wang, 1085–1090. London: Idea Group.