Elektrotehniški vestnik 73(5):285-290, 2006
Electrotechnical Review: Ljubljana, Slovenija

Analyzing the Impact of Investment in Education and R&D on Economic Welfare with Data Mining

Vedrana Vidulin, Matjaž Gams

Jožef Stefan Institute, Department of Intelligent Systems, Jamova 39, 1000 Ljubljana, Slovenia
E-mail: vedrana.vidulin@ijs.si

Abstract. Has greater investment in education and research and development (R&D) a positive impact on economic welfare? We analyzed this question using the Weka machine learning and data mining systems. We collected data from the statistical databases for the year 2001. The obtained classification trees show that the level of participation in higher levels of education has a high impact on economic welfare mainly through more and better educated individuals. Moreover, the indicator of investment in R&D is being assessed as an important promoter of economic welfare. With the resulting classification trees we can with the 70% accuracy predict the income class of the country on the basis of its education and R&D.

Keywords: economic welfare, education, R&D, data mining, classification tree

Vpliv investicij v izobraževanje in R&R na gospodarsko rast

Povzetek. Analizirali smo vprašanje, ali večja vlaganja v izobraževanje ter raziskave in razvoj (R&R) neposredno vplivajo na večjo gospodarsko rast. Pri tem smo uporabili programski paket Weka, tj. sisteme za strojno učenje in rudarjenje podatkov. Analizo smo opravili na statističnih podatkih za leto 2001, ker smo za to leto dobili največ podatkov. Zgrajena klasifikacijska drevesa kažejo, da izobraževanje, posebno visokošolsko izobraževanje, pomembno vpliva na gospodarsko rast. Tudi vlaganja v raziskave in razvoj so se izkazala za zelo pomembna. Dobljena drevesa so ponavadi dosegala 70-odstotno točnost. To pomeni, da na podlagi atributov visokega šolstva in raziskav ter razvoja lahko s 70-odstotno natančnostjo napovemo gospodarski položaj države. Čeprav ta raziskava prikazuje zakonitosti na en način prek strojnega učenja, vseeno močno potrjuje povezavo med omenjenimi področji.
Ključne besede: gospodarska rast, izobraževanje, R&R, rudarjenje podatkov, klasifikacijsko drevo

1 Introduction

The modern society puts a lot of effort in analyzing the level of impact of different social sectors on economic welfare; the sectors that have a greater impact draw more attention on themselves and more money [2, 10]. For the modern society we can also say that it is based on knowledge, therefore education and R&D should and do play an important role in it [4, 12]. But has greater investment in education and R&D actually a positive impact on economic welfare?

Modern approaches to data analyzing using machine learning (ML) and data mining (DM) give us an opportunity to examine this question.

Data mining is a process of analyzing data to identify patterns or relationships [7]. Data mining is about solving problems by analyzing data already present in databases [13].

There is a lot of statistical data available on the Internet for the field of education, R&D and economy covering the period of one year. That gives us an opportunity to apply ML and DM techniques to the existing data seeking for an as accurate classification tree as possible [3]. That tree will be used for predicting the income group of a country on the basis of its education and R&D.

In Section 2 the contents, structure and quality of the statistical data are described together with the DM technique used for answering the question we are interested in. In Section 3 various constructed trees are presented, and Section 4 contains a conclusion and discussion.

2 Description of the Data and DM Technique

None of the statistical databases available on the Internet contains all indicators we are interested in. Therefore, we extracted the data from various statistical databases provided by The World Bank, UNESCO

Prejet 3. april, 2006
Odobren 9. september, 2006

Institute for Statistics, USAID (Global Educational Database) and WIPO Patent Scope.

From the available data we selected 50 different indicators for the year 2001 and exported them in the form of a spreadsheet table. Each country represents a row in the table and each indicator a column.

The final dataset consists of 158 examples (countries), 50 indicators, from which 7 represent economic indicators (e.g. *GDP per capita*, *GDP growth*, *GNI per capita*…), 14 R&D indicators (e.g. *Researchers per million inhabitants*, *GERD per capita*, *Grants of patents*…), 24 educational indicators (e.g. *Tertiary students per 100,000 inhabitants*, *Public expenditure on education as percentage of GDP*, *School life approximation*…) and 5 general indicators (e.g. *Fixed line and mobile phone subscribers per 1000 people*, *Internet users per 1000 people*, *Military expenditure as percentage of GDP*…). All indicators are numeric except one which is discrete.

Discrete indicator *GNI per capita* was chosen for the class. ***Gross National Income*** (GNI) prizes the total value of goods and services produced within a country (i.e. its Gross Domestic Product) together with its income received from other countries (notably interest and dividends) and less similar payments made to other countries [5]. The indicator can take one of the three values, i.e. low, middle or high. This corresponds to the official World Bank classification [6] with intervals:

1) low – $745 or less
2) middle – $746–9,205
3) high – $9,206 or more

The drawback of the final dataset is that in spite of the fact that the countries with a lot of missing data were deleted from the dataset, we still needed to cope with the problem of missing data. That problem is less present in the case of the economic and general indicators.

From the ML and DM techniques available in Weka [13] we chose J48, the implementation of C4.5 [9]. It is a technique used for the induction of classification, i.e. classification trees. In Weka we can graphically represent the obtained tree thus gaining an easy comprehensive answer to the question which indicators have the greatest impact on the income. Also, C4.5 is appropriate in this case because it can deal with the numeric and missing data.

Classification trees are built in a top-down manner. The first task is to choose the appropriate indicator which will be at the root of the classification tree. The next step is to add branches. If a discrete indicator is taken into consideration, than there are as many branches as there are different values of the given indicator. When we use numeric indicators, there are only two branches, one that represents values less or equal than the value on which the numeric indicator was split and one that represents greater values. The set of examples is divided into the number of subsets equal to the number of branches. Examples are distributed into the subsets according to the value of the indicator placed as the root node. Now the process can be repeated recursively for each branch, using only those instances that actually reach the branch. If at any time all instances at a node have the same classification, further branching is stopped [13].

The central choice in the algorithm is selecting which feature to test at each node in the tree. We would like to select the feature that is most useful for classifying examples [1]. Therefore, for every indicator a statistical measure called "information gain" is calculated. Information gain evaluates how well the selected indicator divides the set of examples according to the target attribute, in this case according to the income group of the country.

To calculate information gain, one way is to first calculate entropy. Entropy is a measure of impurity of the set of examples. We calculate entropy using Eq. (1).

$$Entropy\ (S) = \sum_{i=1}^{c} - p_i \log_2 p_i \tag{1}$$

S represents the set of examples for which we are calculating the entropy, c is the number of values of the target indicator, and p_i is the proportion of the examples from the set S belonging to the i-th class.

Using the entropy, information gain is calculated as the expected reduction in entropy caused by distributing the examples according to the chosen indicator. We can calculate information gain using Eq. (2).

$$Gain\ (S, i) \equiv Entropy\ (S) - \sum_{v \in Values\ (i)} \frac{S_v}{S} Entropy\ (S_v) \tag{2}$$

S represents the set of examples for which we calculate information gain, i is the chosen indicator, *Values(i)* address to all possible values of the indicator i, and S_v is the subset of S for which indicator i has value v. Generally, *Gain(S, i)* is the expected reduction in entropy caused by knowing the value of indicator i.

3 Results

We conducted four experiments using different sets of indicators, trying to find as accurate classification trees as possible. In all experiments, the same class was used, i.e. the discrete indicator *GNI per capita*.

In all experiments default values of the classifier parameters set in Weka were used. We experimented with the usage of reduced error pruning, but the obtained classification trees were less accurate. To estimate the accuracy of the trees, we used 10-fold cross-validation [8].

3.1 Experiment with 15 Most Related Indicators

The first classification tree was induced on the basis of indicators that describe investment at all levels of education (e.g. *Total educational expenditure per pupil as a percentage of GDP per capita*), higher education (e.g. *Tertiary students per 100,000 inhabitants*) and R&D (e.g. *Researchers per million inhabitants*). In total, 15 indicators were chosen.

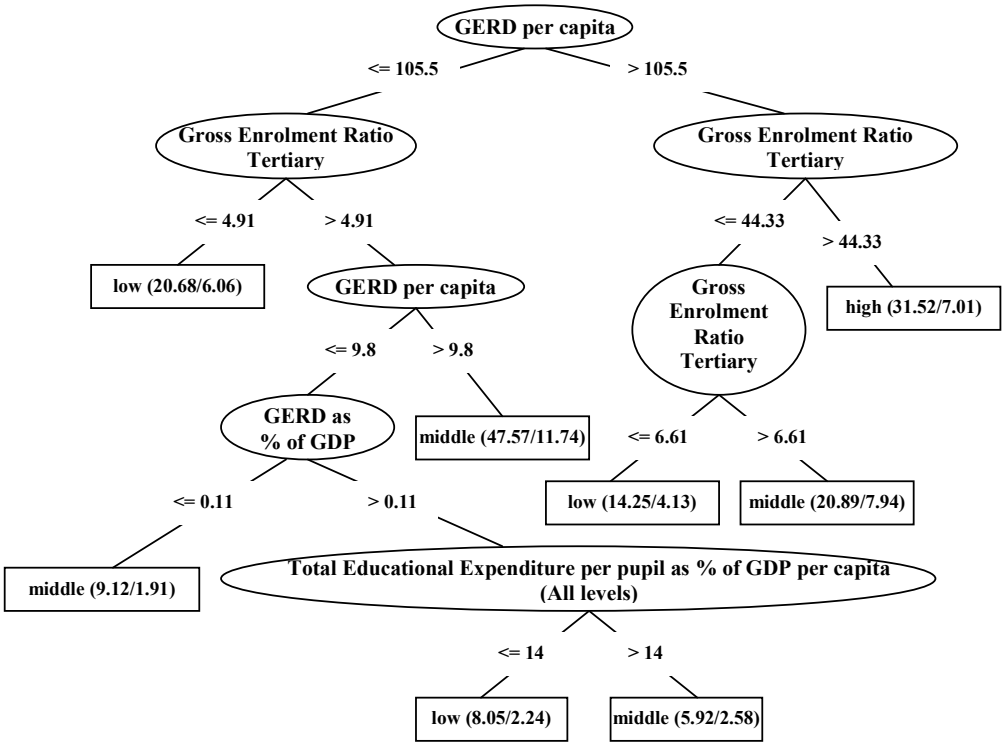

Figure 1. Classification tree with the accuracy of 69.62% built from 15 most related indicators (the number of correctly and incorrectly classified instances is presented at the leaves of the classification tree next to the labels of the class)

Some indicators demand clarification, like *GERD*. It is an abbreviation for ***Gross Domestic Expenditure on R&D*** and denotes expenditure on R&D performed on the national territory during a given period. Calculation also includes R&D performed within a country and funded from abroad and excludes payments made from abroad for R&D. We used value stated per capita in PPP$ and GERD as a percentage of GDP. ***PPP$*** means purchasing power parity stated in American dollars. PPP are the rates of currency conversion that allow for differences in price levels between countries. Normally they are given in national currency units per US dollar [12].

Gross enrolment ratio is enrolment at a given level of education, regardless of age, expressed as a percentage of the population in the theoretical school-age group corresponding to this level of education. Gross Enrolment Ratio is widely used to show the general level of participation in a given level of education. It indicates the capacity of the education system to enrol students of a particular age-group [11].

Total educational expenditure per pupil as a percentage of GDP per capita presents the part of the GDP per capita that is spent on each student, what indicates the level of investment in the development of the human capital.

From the induced tree presented in Fig. 1 several interesting conclusions can be drawn. Firstly, there are two distinctive groups of countries. Those with high investment in R&D and tertiary education have also high GNI, while those with low investment in both have low GNI. Some other relations are more complex. For example those countries that invest more in R&D, but have a very low level of enrolment in the tertiary education, fall into the low-income group (see the right subpart of the tree). But those with at least reasonable investment in the tertiary education and low R&D still fall into the middle-income group (left subpart).

The accuracy of the obtained tree is 69.62%, which can be regarded as reasonable.

3.2 Experiment with 34 Indicators

The idea behind the second experiment was to enlarge the number of indicators, so we chose 34 of them. In addition to indicators that describe higher education and R&D, we included some general indicators (e.g.

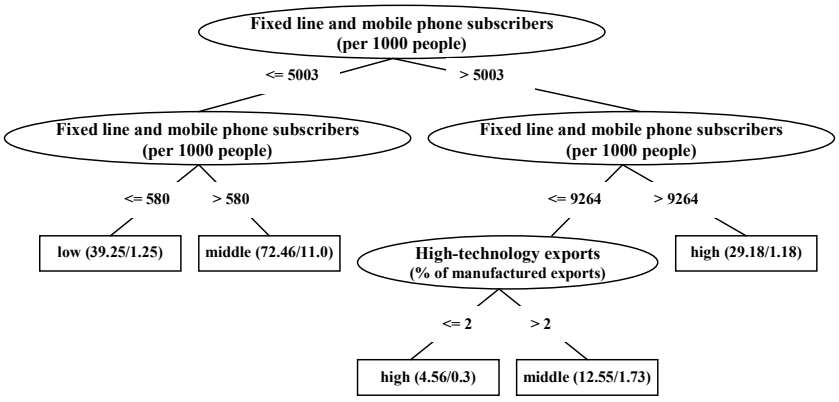

Figure 2. Classification tree obtained from economic, education and R&D indicators and with the accuracy 82.91%

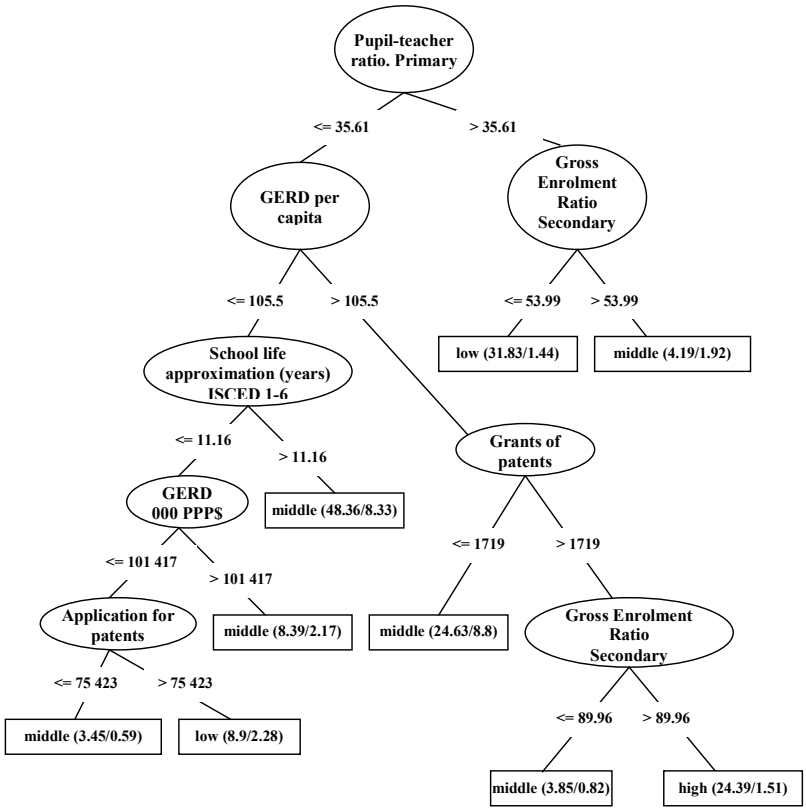

Figure 3. Classification tree obtained from 39 indicators and the resulting accuracy of 70.89%

Personal computers per 1000 people) and some economic indicators (e.g. *Exports of goods and services*).

The constructed tree is presented in Fig. 2. The indicator *Fixed line and mobile phone subscribers* plays the most important role in this tree because it directly shows the wealth of a country.

From the indicators that we are interested in, only ***High-technology exports*** is present in the tree structure. This indicator represents the percentage of the high-technology exports in the overall export. However, some countries that have lower high-technology exports belong to the high-income group. An insight into the data showed that these are the countries rich in oil and natural gas, therefore these are the primary promoters of their economic welfare.

Although the constructed tree has a relatively high accuracy, i.e. 82.91%, the presented indicators are not in the focus of our interest. They do not represent an answer to our question. Furthermore, some of the relations are too direct in the sense that "richer countries have more money".

3.3 Experiment with 39 Indicators

The third experiment was conduced with the purpose to get a better insight into the role of education. In this case we chose indicators that describe all levels of education (e.g. *School life approximation (years). ISCED 1-6*), together with the indicators of R&D. In total, 39 indicators were chosen.

Indicator ***Pupil-teacher ratio*** denotes the average number of students per teacher at a specific level of education in a given school-year.

School life approximation for the ISCED levels 1-6 denotes the number of years a child of the school entrance age is expected to spend at school, or university, including years spent on repetition. This indicator shows the overall level of development of an educational system in terms of the number of years of education that a child can expect to achieve [11].

The obtained tree is presented in Fig. 3. What characterizes the high-income countries are the high quality of primary education (classes with fewer pupils), high level of enrolment in secondary education, high investment in R&D and more granted patents than in the middle and low-income countries. On the other hand, in the low-income countries children abandon school at an early age, e.g. after finishing the elementary school. Therefore, the level of enrollment in the secondary education is lower than in the middle and high-income countries. The value on which the *School life approximation* indicator splits values also confirms that claim. In the low-income countries it is expected that children who are now entering the educational system spend less or equal than 11.16 years in the school including years spent on repetition. Some countries that belong to the middle-income group also have the same lower school life expectancy, but in most cases they invest more money in R&D. The most successful countries invest much in primary and secondary school and R&D and have a high number of granted patents.

The accuracy of the obtained tree is 70.89%. Since we are more interested in the higher education and globally in investment in education, the next experiment will not take into account indicators from the lower levels of education.

3.4 Experiment with 26 Indicators

In this experiment we concentrated on higher education and R&D and chose 26 indicators.

The constructed tree is presented in Fig. 4. From the tree we can clearly see that high-income countries invest more in R&D, tertiary education, have more granted patents and higher percentage of high-technology exports. Low-income countries, on the other hand, differ. Most of them have a lower number of granted patents and low school-life approximation. School-life approximation is similar to the one from the third experiment (less than or equal to 11.16 in the third experiment and less than or equal to 11.22 in this experiment). This tree shows that education has a positive impact on economic welfare. In both cases the split on the educational indicator the less or equal side of the tree guides to the low-income class.

The situation is a bit less clear with R&D. Countries with a higher number of granted patents and higher investment in R&D can still belong to the low-income group. Among the countries in the middle-income group there exist some differences. They may have more or less granted patents or higher or lower level of investment in R&D, but when we take education into account, they are all on the right (greater than) side of the tree.

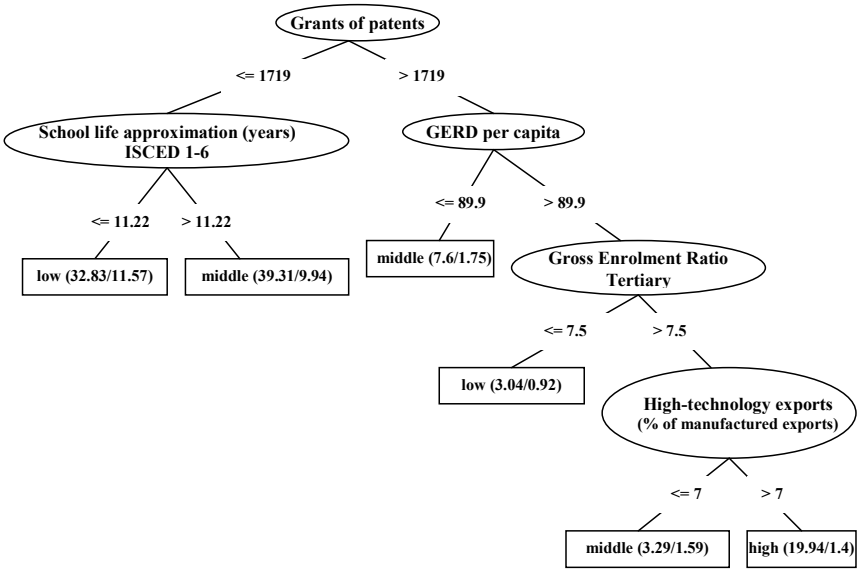

Figure 4. One of the most informative trees obtained from 26 most related indicators and the resulting accuracy of 70.25%

The obtained tree is 70.25% accurate. From the conducted experiments it can be concluded that in all the cases, except in the second experiment, the obtained level of accuracy is around 70%.

4 Conclusion

In our investigation we use the ML and DM techniques for answering the question if greater investment in education and R&D has a positive impact on economic welfare. Besides indicators that show the amount of invested money, the indicators that describe education and R&D were employed too.

The presented results show a minor importance of high investment in low-level education in contrast to the importance of high level of participation in higher levels of education. This can be seen from the fact that countries with the low percentage of people enrolled in secondary and tertiary education belong to the low income group. Indicators that describe just the amount of money invested in education do not appear essential, but the importance of education is obvious since the countries that do invest in R&D while their level of enrolment in tertiary education is low belong to the low-income group.

In the case of R&D, besides investment in it, the number of granted patents and amount of high-technology exports also play an important role.

From the obtained results we can conclude that investing in R&D has an important impact on economical welfare, provided that people are oriented towards high technologies thus assuring the progress of the economy to be satisfactory.

We propose the classification tree obtained with the fourth experiment to be used for predicting the income group of a particular country on the basis of indicators of education and R&D. The accuracy of such prediction is approximately 70%.

Although the obtained trees give us only a partial idea of the state and are still far from being regarded as a final proof, the positive impact of high-level knowledge (tertiary education and R&D) on economic welfare can be clearly seen.

5 References

[1] I.S.I Abuhaiba, Arabic Font Recognition Using Decision Trees Built from Common Words, *Journal of Computing and Information Technology – CIT*, 13, 3, p.p. 211–223, 2005

[2] B. Baumohl, *The Secrets of Economic Indicators: Hidden Clues to Future Economic Trends and Investment Opportunities*, Wharton School Publishing, 2004

[3] Carnegie Mellon Professor Chosen to Head Google Engineering Office in Pittsburgh, http://www.cmu.edu/PR/releases05/051215_moore.html

[4] Globalization - Economic Growth and Development and development indicators, http://www.planetpapers.com/Assets/4302.php

[5] Gross National Income – Wikipedia, 2006, http://en.wikipedia.org/wiki/Gross_National_Income

[6] *Human Development Report 2003 – Millennium Development Goals: A compact among nations to end human poverty*, pp. 362, Oxford University Press, 2003

[7] Iomega NAS Terms Glossary, 2005, https://iomega-eu-en.custhelp.com/cgi-bin/iomega_eu_en.cfg/php/enduser/std_adp.php?p_faqid=1725

[8] R. Kohavi, A Study of Cross-Validation and Bootstrap for Accuracy Estimation and Model Selection, *IJCAI*, p.p. 1137-1145, 1995

[9] J.R. Quinlan, *C4.5: Programs for Machine Learning*, Morgan Kaufmann, 1993

[10] B. Rusjan, E. Rusjan, Fizikalni pogled na nihanja gospodarstva in financ, *Elektrotehniški vestnik*, 73, 1 38-44, 2006

[11] UNESCO Institute for Statistics, 2002, http://www.uis.unesco.org/ev_en.php?ID=2867_201&ID2=DO_TOPIC

[12] D.N. Weil, *Economic Growth*, Addison-Wesley, 2005

[13] I.H. Witten, E. Frank, *Data Mining – Practical Machine Learning Tools and Techniques*, Chapter 4: Algorithms: The basic methods, pp. 97-105, Elsevier Inc., 2005

Vedrana Vidulin received her university degree in 2005 from the Faculty of Philosophy, University of Rijeka, Croatia, by defending her thesis "Neural Networks: Algorithms and Applications in Education". She is continuing her education at the Jožef Stefan International Postgraduate School, study program New Media and e-Science, and is receiving scholarship from the Jožef Stefan Institute, Ljubljana, Slovenia.

Matjaž Gams is an Associate Professor of computer and information science at the University of Ljubljana and a Senior Researcher at the Jožef Stefan Institute, Ljubljana, Slovenia. He teaches several courses in computer science at graduate and postgraduate levels at Faculties of Computer and Information Science, Economics, etc. His research interests include artificial intelligence, intelligent systems, intelligent agents, machine learning, cognitive sciences, and information society. In his publication list there are over 300 items, 50 of them in scientific journals. He has headed several major artificial intelligence applications in Slovenia, including the major national employment agent on the Internet, and the Slovenian text-to-speech system donated to several thousand users.