INTRODUCTION TO EDUCATIONAL DATA MINING USING MATLAB

Lukáš Zoubek

Department of Information and Communication Technologies,

Pedagogical Faculty, University of Ostrava

Abstract

Identification of significant differences in sets of data represents frequent data mining application. This paper presents initial study introducing application of MATLAB in the area of educational data mining tasks. In the concrete, statistical test called Analysis of Variance is used to determine significant difference of data sets analyzed. The study has been performed on real data coming from pedagogical tests focused on evaluation of mathematical skills of secondary school students in Czech Republic.

1 Introduction

In the modern world, information and communication technologies become to be fully involved in the educational processes. As the result, large amount of data can be generated as a side effect. The data can be either directly related to the educational process (e.g. testing of students, learning management systems LMS) or can be generated from other related activities (e.g. leisure time activities, library entries). Educational data may contain many interesting information and potential knowledge about students and their learning habits. However, such knowledge is hidden and their extraction is not a trivial task.

Many applications try to describe and then effectively apply the knowledge hidden in the databases or in large data warehouses. Knowledge acquiring methods can be used to extract non-trivial, previously unknown and potentially effective and useful information (knowledge) from the data. The principle of these methods is mainly based on the application of analytic methods. They typically use specially preprocessed data as an input and return knowledge information as an output. This branch of science is also called Data mining, Information harvesting or Knowledge discovery in databases.

The objective of this paper is to present capability of MATLAB as a data mining tool when applied on data coming from educational tests of secondary school students. In the concrete, application of techniques for identification of statistically significant differences among data sets is described. The study has been realized at Department of information and communication technologies (University of Ostrava) which focuses its research activities on extraction of knowledge from educational data.

2 Original data

For the need of the study, set of data coming from research realized at more than 90 secondary schools in the Czech Republic has been used. All the schools are located in Moravia-Silesian region. In the total, about 8 000 students were tested in four subjects (mathematics, native language (Czech), foreign language (English or German) and general study pre-requisites) [1].

The secondary schools engaged in the research are subdivided into nine groups depending on their specialization – factor *Type of school*. The factor levels are as follows:

- Economic *(ECO)*,
- Grammar school - gymnasium *(GRA)*,
- Lyceum *(LYC)*,
- Social and health studies *(SAH)*,
- Natural science *(NAT)*,
- Trade and service *(TAS)*,

- Social science *(SOC)*,
- Technical *(TEC)*,
- Art studies *(ART)*.

Only the mathematical skills have been selected out of the four subjects tested during the original research. To evaluate mathematical skills, each student had to answer 61 mathematical questions during the original research. The correctness of each answer has been then encoded into a binary value. The correct answer is represented by value 1, while the wrong answer is represented by value 0.

Table 1: NUMBER OF STUDENTS DEPENDING ON THE TYPE OF SCHOOL

Type of school	*ECO*	*GRA*	*LYC*	*SAH*	*NAT*	*TAS*	*SOC*	*TEC*	*ART*	Total
Number of students	734	2 086	800	633	245	937	109	2 284	78	7 906

These 61 various mathematically oriented questions have been specially prepared in cooperation with pedagogical experts so as to cover eight mathematical skills – factor *Mathematical skill*. The corresponding factor levels can be characterized as follows:

- Understanding of the number as a concept expressing quantity (*skill1*);
- Numerical skills (*skill2*);
- Understanding of mathematical symbols and signs (*skill3*);
- Orientation and work with table (*skill4*);
- Graphical reception and work with graph (*skill5*);
- Understanding of plane figures and work with them, spatial imagination (*skill6*);
- Function as a relation between quantities (*skill7*);
- Logical reasoning (*skill8*).

Then, in the next step of data preparation, each of the eight mathematical skills presented above has been evaluated depending on the corresponding answers. For each student, the skills have been evaluated separately. Each of the skills has been characterized by a percentage (0-100) representing the level of the individual skill. The evaluation strategy has been prepared again in cooperation with pedagogical experts. So, at the end, each student has been represented by a vector of eight values corresponding to eight skills (attributes).

After cleanup, data about 7 906 students (males and females together) have been obtained. Table 1 shows distribution of students depending on the type of secondary school. So, the final database contains 7 906 recordings, where each recording is represented by nine attributes. 8 attributes characterize levels of the 8 mathematical skills and the last attribute represents type of school.

3 The method

The research presented in this paper comes from the initial analysis of the database realized at the Department of information and communication technologies before. In the paper [2], authors present analysis of the database using statistical tests of hypotheses. In the concrete, Student's t-test for testing the equality of means has been used. To visualize relations obtained by circular permutation of the statistical tests, Haase diagrams have been proposed in the paper. The drawback of the proposed technique is that large amount of simultaneous statistical tests performed increases the test error far beyond the significance level α [3]. So, the inequalities obtained should be considered only as hypotheses indicating some interesting relationship within data.

To extend the initial research, *Analysis of Variance* (ANOVA) methods should be used instead of *Two-Sample* test. MATLAB and its Statistics toolbox contain various tools performing Analysis of Variance [4]. The aim of the ANOVA analysis in the project is to compare differences among levels of the skills obtained for various groups of students.

One-way ANOVA

Analysis of Variance is a statistical test used to determine whether three or more data sets (means) are statistically significantly different. In the present study, one-way ANOVA has been used as an initial technique to analyze significant differences between either types of schools or various mathematical skills analyzed.

Selection of proper ANOVA test depends on the data to be analyzed. First criterion affecting selection of ANOVA test corresponds to distribution of the data sets. Because the data analyzed do not follow normal distribution (Lilliefors test), nonparametric version of the classical one-way ANOVA should be used. There are two suitable tests implemented in MATLAB that can be used (Kruskal-Wallis test and Friedman's test). Both the tests analyze the ranks of the data rather than their original numeric values (skill levels). Ranks are found by ordering the data from smallest to largest across all groups, and taking the numeric index of this ordering.

Kruskal-Wallis test is a nonparametric test that compares three or more unpaired groups of data. Moreover, MATLAB implementation of the test allows analysis of data where number of recordings is not evenly distributed into individual groups to be analyzed (unbalanced model). It is a case of *Type of school* factor as could be seen in Table 1. The Kruskal-Wallis test evaluates the hypothesis that all samples come from populations that have the same median, against the alternative that the medians are not all the same. As a result, *p*-value for the null hypothesis that all samples are drawn from the same population is obtained [4].

Friedman's test is a nonparametric test suitable for analysis of paired groups of data. Paired groups correspond for example to different treatments or repeated measures. In our data set, *Mathematical skill* factor can be considered as paired data type. The Friedman test ranks the values in each row representing individual student separately.

To provide advanced visual representation of the results obtained during initial ANOVA analysis, a multiple comparison test (*multcompare*) should be used to interactively determine which pairs of data sets are significantly different. Using this test, concrete groups (types of school or mathematical skills) can be easily characterized and compared, and general results can be then determined.

N-way ANOVA

N-way analysis of variance is a statistical test used to evaluate possible interaction between the factors (variables) characterizing the data. Interaction can be understood as combined effect of the analyzed factors on the dependent measure (*level of the skill*). Since the set of data is characterized as unbalanced, *ANOVAN* function has been used instead of classical two-way ANOVA. As a result of *ANOVAN* function, *p*-values for null hypotheses on the main effects of analyzed factors and interactions at all levels can be computed.

4 Results

This section of the paper summarizes results obtained when the real educational data presented above have been analyzed using MATLAB software. In the first step, effects of individual factors have been analyzed. As a final analysis, *N*-way analysis of variance has been performed in order to evaluate also interaction between the two factors analyzed (*Type of school* and *Mathematical skill*).

Effect of Mathematical skill factor

In the first part, differences in skill levels depending on various mathematical skills analyzed have been evaluated. Table 2 presents order of the skills depending on the average skill level achieved. The aim of the analysis is to determine whether skill levels achieved are significantly different for various skills to be analyzed.

Table 2: AVERAGE SKILL LEVELS COMPUTED FOR SKILLS ANALYZED

Skill	Average skill level
skill1, skill4	70%
skill2	68%
skill3, skill8	57%
skill7	54%
skill6	50%
skill5	42%

Now, the information in Table 2 can be compared to the result of Friedman's test followed by multiple comparison test. Friedman's test (p-value) indicates that there is a significant difference in the levels obtained for various mathematical skills. The final graph summarizing multiple comparison test is presented in Figure 1. Two means are significantly different if their intervals are disjoint (e.g. *skill5* and *skill6*), and are not significantly different if their intervals overlap (e.g. *skill1* and *skill2*).

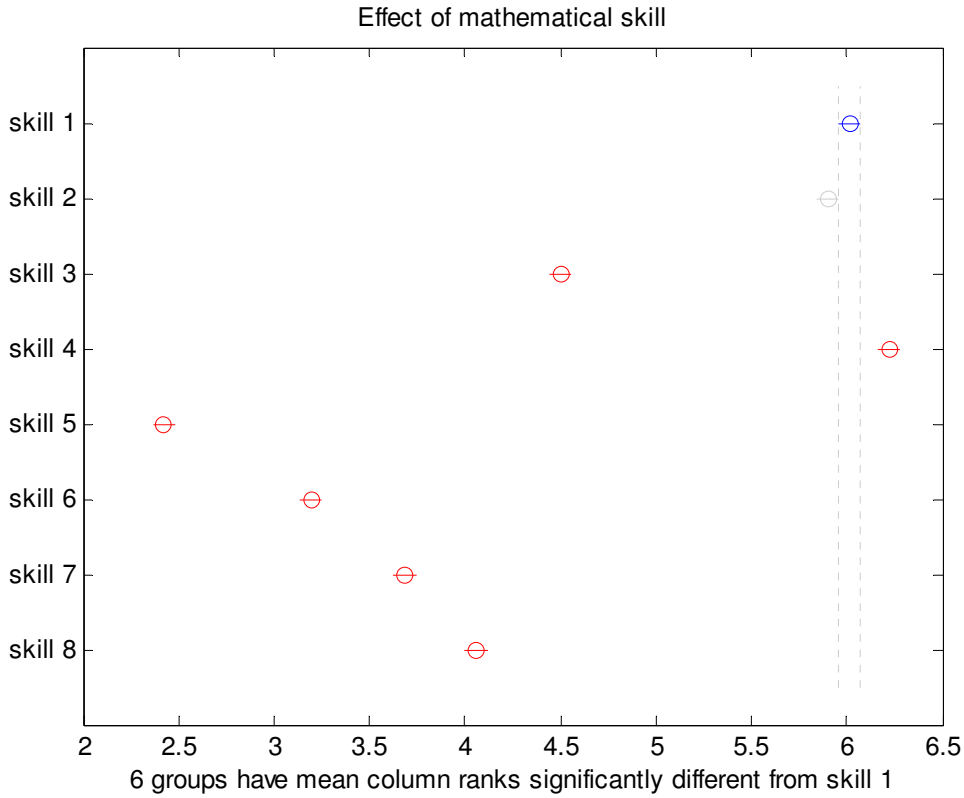

Figure 1: Graph visualizing effect of *mathematical skill* factor

As could be seen, *skill4* is characterized by the highest skill level (median of ranks) which is also significantly different from the other skills. Multiple comparison test also indicates that *skill1* and *skill2* are not significantly different each other. Figure 1 also shows that the students reached significantly lowest skill level for *skill5*.

Similar results have been obtained when skill effect has been analyzed for the individual types of school separately. The main difference is that levels obtained for *skill1*, *skill2* and *skill4* are not significantly different for most of the secondary school types. Only for Social and health study schools (*SAH*) the level for *skill4* is significantly different – higher. There is also difference in the absolute values of average skill levels obtained for different types of school.

Effect of Type of school factor

As could be supposed, average skill levels vary depending on the type of school. So, detailed analysis of *Type of school* factor has been performed so as to evaluate effect of the school type on skill levels achieved.

Firstly, analysis of the *Type of the school* factor has been done without separation of the individual analyzed skills. Kruskal-Wallis test again indicates significant difference in the levels obtained for various secondary schools. The overall characteristic can be seen in Figure 2. Grammar schools (*GRA*) are characterized by significantly highest average level of analyzed skills. Technical schools (*TEC*) and Lyceums (*LYC*) reached second highest average level, significantly different from all other schools. There is no simple rule to characterize order of the remaining secondary schools.

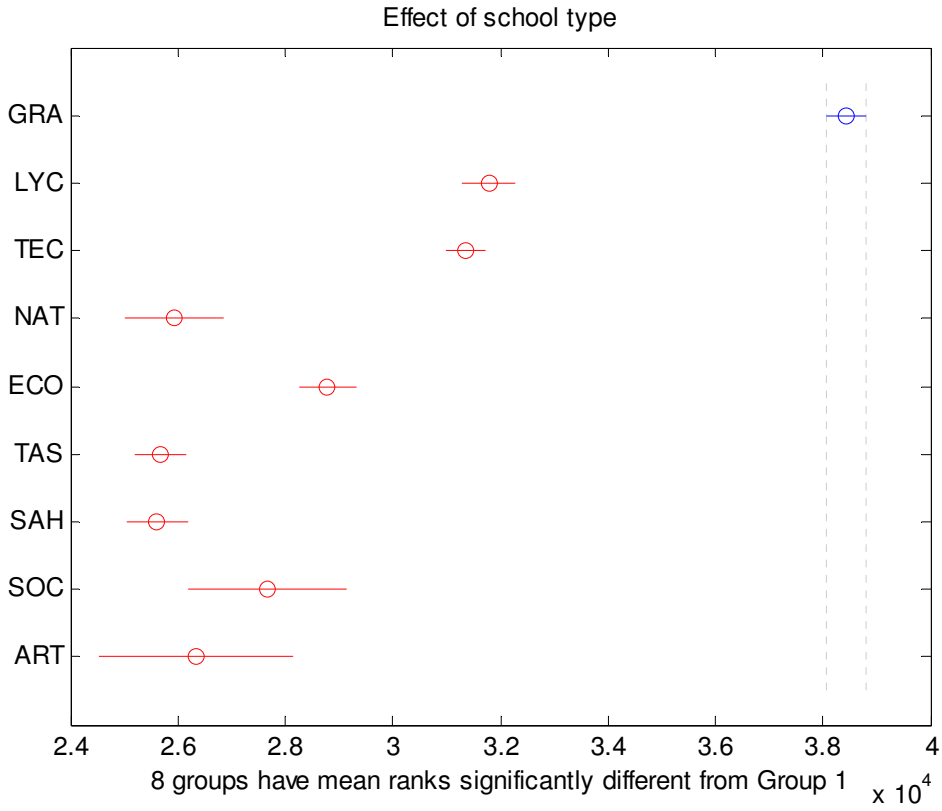

Figure 2: Graph visualizing effect of *type of school* factor

Again, similar results have been obtained, when individual skills have been considered separately. Grammar schools are characterized with the highest average level achieved by the students. Grammar schools are typically followed by Lyceum and Technical school students. Order of the other secondary schools slightly varies depending on the concrete mathematical skill. The main difference is in the case of the *skill5*.

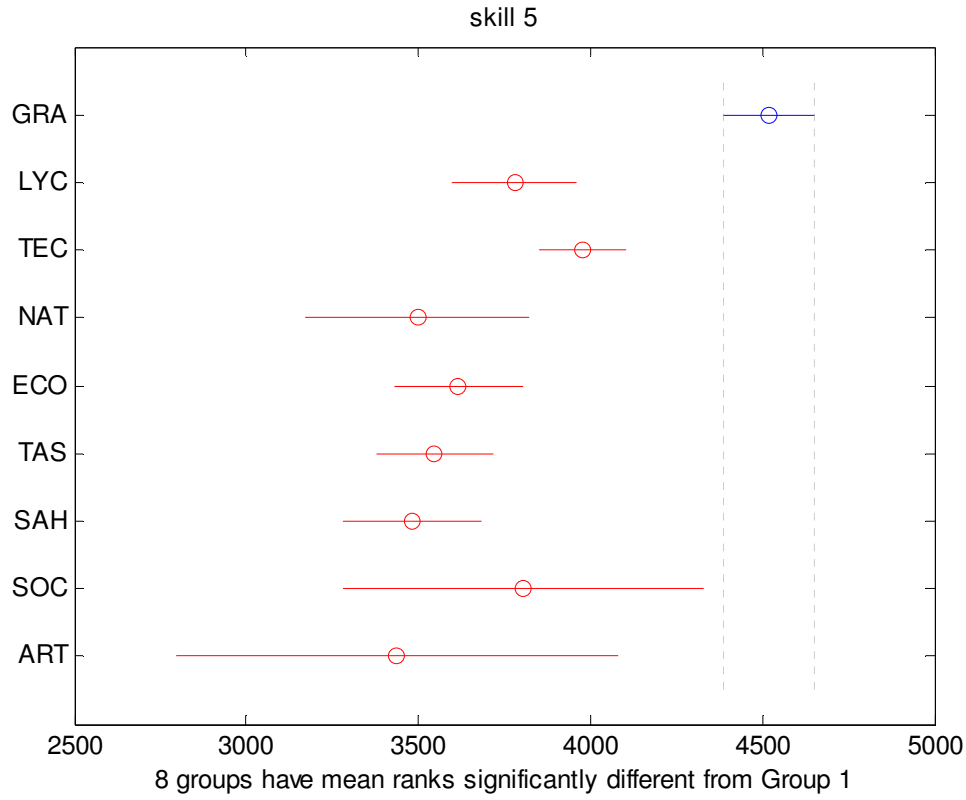

Figure 3: Graph visualizing effect of *type of school* factor obtained for *skill5*

In the case of the *skill5*, the difference between the highest average skill level (*GRA*) and the lowest average level is small. The difference is only 8.5% which is the smallest value among all analyzed mathematical skills. This fact corresponds to the analysis of the individual skills, where average skill level achieved for *skill5* has been significantly the lowest for all secondary schools. Figure 3 shows results of the analysis performed for *skill5*. Grammar schools (*GRA*) again achieved the highest average level, but in this case are closely followed by remaining schools which are hard to be separated each other.

Interaction of both the factors

Previous analyses characterized partial effects of analyzed factors on the data. To summarize the results, two statements can be derived. Type of school affects the level of skills achieved by the students. Moreover, level of the skills depends on the mathematical skill to be tested. Now, the question is: *Is there a synergistic effect of type of school and analyzed skill on the obtained skill level*?

The answer has been obtained using *N*-way ANOVA analysis. The result (*p*-value) proved interaction between the two factors analyzed in the research. It means that effect of the type of school depends on the mathematical skill. Moreover, *N*-way analysis confirmed also main effects of the factors.

5 Conclusions

Analysis of variance is a powerful tool suitable for analysis of more than two data sets. This paper shortly presents application of such a statistical test on real educational data characterizing mathematical skills of secondary school students. During the research, two main factors describing the data have been analyzed separately using one-way ANOVA and then interaction of the factors has been evaluated using *N*-way analysis.

The paper also characterizes some of research activities realized at Department of information and communication technologies and presents application of MATLAB software in educational data mining.

References

[1] L. Kubincová, M. Malčík. *Testing of skills of the 1st year secondary schools pupils*. In: Information and Communication Technology in Education, Rožnov pod Radhoštěm, Czech Republic, 2008.

[2] L. Zoubek, M. Burda. *Visualization of Differences in Data Measuring Mathematical Skills*. Cordoba, Spain, 2009. ISBN 978-84-613-2308-1

[3] R. G. Miller. *Simultaneous statistical inference*, 2nd edition. Springer, 1981. ISBN 978-0387905488.

[4] *Statistics Toolbox User's Guide*. (September 2009), Available at: <http://www.mathworks.com/access/helpdesk/help/pdf_doc/stats/stats.pdf>

Ing. Lukáš Zoubek, Ph.D.
Department of Information and Communication Technologies
Pedagogical faculty
University of Ostrava
Českobratrská 16
701 03 Ostrava
tel: +420 597 092 632
e-mail: lukas.zoubek@osu.cz