Visualization of Differences in Data Measuring Mathematical Skills

Lukáš Zoubek, Michal Burda
{Lukas.Zoubek, Michal.Burda}@osu.cz
Department of Information and Communication Technologies, Pedagogical Faculty, University of Ostrava, Českobratrská 16, 701 03 Ostrava, Czech Republic

Abstract. Identification of significant differences in sets of data is a common task of data mining. This paper describes a novel visualization technique that allows the user to interactively explore and analyze differences in mean values of analyzed attributes. Statistical tests of hypotheses are used to identify the significant differences and the results are then presented using Hasse diagrams. The presented technique has been tested on real data coming from pedagogical tests focused on evaluation of mathematical skills of secondary school students in Czech Republic. The results show that the proposed tool provides comprehensible representation of the data.

1 Introduction

Knowledge discovery from databases (also known as Data Mining) is a methodology for extraction of non-trivial, previously unknown, and potentially useful knowledge from data [4]. It is broadly used in a commercial sector, research and other domains. A characteristic feature of Data Mining methods is an intensive utilization of computers for difficult computations and testing of large amount of combinations.

The objective of this paper is to present the results of application of a data mining method on data coming from educational tests of secondary school students. In the concrete, a technique for identification of statistically significant differences among mean values is described.

Such method together with the novel visualization technique described here allows the analyst to explore data and view significant differences among mean values of groups of students. The process is on-line: the attributes used to partition the data into groups are set interactively by the user. The results are immediately presented in a graphical form and the user is allowed to change settings in order to allow him or her to iteratively explore the data and find some useful knowledge.

1.1. Related work

An extensive amount of research has been done on data exploration and data mining. Let us focus on visualization techniques related to the main objective of this paper only.

Eick in [3] presents three interesting techniques, where 3D bar chart, scatterplot and a combination of para-boxes, bubble plots and box plots allow to visually analyze values of quantitative attributes.

Authors of [5] describe a visualization of hypothesis tests in multivariate linear models by representing hypothesis and error matrices of sums of squares and cross-products as ellipses, implemented for R, an open-source statistical software [10].

Two prevailing approaches to visualize association rules [1] are compared in [11]. First approach uses two-dimensional matrix to view support and confidence of the rules. Another approach is to use directed graph. The nodes of the graph represent items, and the edges represent the associations. Paper [6] experiments further with animation of the edges to depict the associations.

The co-author of this paper has discussed concept lattices and the approach that utilizes Hasse diagram with negative edges. In [2], these two techniques are compared.

2 Original data

To evaluate performance of the presented analytic tool, a database consisting of educational data has been used. The database comes from research realized at more than 90 secondary schools in the Czech Republic. All the schools are located in Moravia-Silesian region. During the original research, about 8000 students were tested in mathematics, native language (Czech), foreign language (English or German) and general study pre-requisites [7].

The secondary schools engaged in the research can be split into nine categories depending on their orientation and specialization. The categories 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)*.

Another data attributes about the students are sex, age, and city. 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 the school.

For the need of our actual research presented in the article, only the mathematical skills have been analyzed. During realization of the original research, each student had to answer 61 mathematical questions. 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 and sex

Type of school	Number of males	Number of females
ECO	212	522
GRA	807	1 279
LYC	309	491
SAH	47	589
NAT	102	143
TAS	224	713
SOC	8	101
TEC	1 965	319
ART	18	60
TOTAL	**3 692**	**4 214**

The test questions have been specially prepared in cooperation with pedagogical experts so as to cover eight important mathematical skills. They 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*).

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 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).

3 The method

On the above described data, a method for searching statistically significant differences among mean values has been applied. We have been searching for significant differences among the means (averages) of mathematical skills.

To identify significant differences, a statistical test of hypotheses could be used. For our purpose, a two sample Student's t-test for testing the equality of means has been used [9]. The test statistic is:

$$t = \frac{\overline{X} - \overline{Y}}{S}, \quad \text{where} \quad S = \sqrt{\frac{s_X^2}{m} + \frac{s_Y^2}{n}},$$

and where $\overline{X}$ and $\overline{Y}$ are the means of the two samples, $s^2{}_X$ and $s^2{}_Y$ are the sample variances

$$s_X^2 = \frac{1}{m-1}\sum_{i=1}^{m}(\overline{X} - x_i)^2 \quad \text{and} \quad s_Y^2 = \frac{1}{n-1}\sum_{i=1}^{n}(\overline{Y} - y_i)^2 .$$

The test statistic t has Student's distribution with

$$f = \frac{(s_X^2 / m + s_Y^2 / n)^2}{(s_X^2 / m)^2 / (m-1) + (s_Y^2 / n)^2 / (n-1)}$$

degrees of freedom. Thus, for sufficiently high $|t|$, say $|t| > T_f(1\text{-}0.05)$, where T_f is a cumulative distribution function of the Student's distribution with f degrees of freedom, we can reject the hypothesis of equal means, that is, we can consider $\overline{X}$ and $\overline{Y}$ to be statistically significantly different.

This way we can test each combination of mean values. Consider e.g. data in the following table:

Table 2. Table shows aggregated data representing *skill1*. (Variance is a square of stdev)

	ART	ECO	GRA	LYC	NAT	SAH	SOC	TAS	TEC
average	66.02	66.5	77.24	70.37	62.32	62.66	65.83	63.03	69.59
stdev	16.52	16.55	14.16	15.96	15.59	16.5	15.83	17.1	16.52
count	78	734	2086	800	245	633	109	937	2284

By testing each pair of the mean values, we can obtain the following inequalities that represent statistically significant differences:

ART < GRA; ART < LYC; NAT < ART; SAH < ART; ART < TEC; ECO < GRA; ECO < LYC;
NAT < ECO; SAH < ECO; TAS < ECO; ECO < TAC; LYC < GRA; NAT < GRA; SAH < GRA;
SAH < GRA; SOC < GRA; TAS < GRA; TEC < GRA; NAT < LYC; SAH < LYC; SOC < LYC;
TAS < LYC; NAT < SOC; NAT < TEC; SAH < SOC; SAH < TEC; SOC < TEC; TAS < TEC.

Generally, the described technique proceeds as follows:

1. A test characteristic c is selected, i.e. the attribute whose average differences we would like to explore (e.g. some mathematical skill, in our case).

2. Optionally, a selection condition is defined. Selection condition determines, which data rows will be processed only (e.g. grammar schools only).

3. A partitioning attribute is selected (e.g. sex). The partitioning attribute is a categorical attribute that is used to partition the data into groups G_1, G_2, …, G_n, among which the differences of means would be analyzed.

4. A statistical testing of differences among *c*'s mean values of groups G_1, G_2, …, G_n is performed. That is, the difference of mean values among all combinations of groups G_i and G_j are tested. We have used two-sample Student's t-test with level of significance $\alpha = 0.05$.

5. As the result, a relation describing statistically significant inequalities among the groups is obtained: $G_i > G_j$ with respect to *c*.

Thus, the obtained inequalities are based on statistical testing of hypotheses. The results may be very interesting to the analyst. Unfortunately, plain textual representation of the obtained relationships seems not to be very synoptic. *Is there any way of representing them graphically?*

The obtained inequalities may be visualized using a Hasse diagram. Hasse diagram is a graph with each group G_i being represented with a vertex. A downward line is drawn from G_i to G_j, if the statistical test has indicated that the mean value computed for group G_i is significantly greater than mean value computed for G_j (i.e. $G_i > G_j$) and there is no such G_k that $G_i > G_k$ and $G_k > G_j$.

Generally, the Hasse diagram should be understood as follows: a node X is significantly greater than Y, if there exists a downward path from X to Y. The path from X to Y may lead through other nodes – however, it must be always downward. Thickness of the line represents intensity of the difference.

For instance, see Figure 2 depicting inequalities extracted from example data characterized in Table 2. From Figure 2 can be for example seen, that grammar schools (GRA) have the greatest average skill level, whereas natural science (NAT) and social and health studies (SAH) are the worst, but there is not a significant difference among them. Similarly, lyceum (LYC) and technical schools (TEC) are not different either.

Please note that accordingly to the theory of statistics, performing large amount of simultaneous statistical tests increases the test error far beyond the level of significance α [8]. Therefore, the obtained inequalities should be considered only as hypotheses indicating some interesting relationship within data – we can never treat the results as a sure and proven knowledge, if obtained that way.

4 Results

This section presents the results of the proposed tool when applied to a set of real data. The data characterizing mathematical skills of secondary school students have been analyzed from three points of view.

4.1. Male or female

The aim of the first test is to analyze difference between male and female students over the eight mathematical skills analyzed. In the first part, the type of secondary school has not been considered for the test. The results show significant differences in average values of levels for all analyzed skills. For all skills, the average values computed for male students are significantly higher. Hasse diagram characterizing this situation is shown in Figure 1. The lowest difference (2.79%) is obtained for *skill4* (males 71.88%; females 69.09%). On the contrary, the maximal difference (5.57%) between males and females is in the case of *skill6* (males 53.49%; females 47.92%).

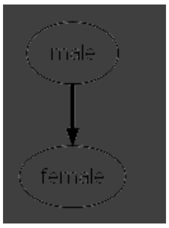

Figure 1. Hasse diagram representing the situation, when the average level computed for male students is significantly different compared to female students

The results of the detailed analysis, when the different types of secondary school have been separated, show, that the secondary schools could be sorted into three groups. Grammar schools (GRA), lyceums (LYC) and economic schools (ECO) can be characterized by the fact that the average skill level characterizing all analyzed skills is significantly higher for male students. In the case of natural science (NAT), trade and service (TAS), social and health studies (SAH), and technical schools (TEC), only for some skills is the average level computed for males significantly higher than for females. The concrete skills and types of school are summarized in the Table 3. The results for remaining schools (art studies (ART), social science (SOC)) do not show significant difference of average skill level for any skill. Unfortunately, relevancy of the data characterizing male students at social science secondary school is low because of very small number of recordings (only eight male students).

Table 3. In the case of four schools, only several skills show significant difference of average skill levels

Type of school	Significantly different skills
NAT	*skill1, skill2, skill3, skill6, skill7, skill8*
TAS	*skill1, skill2, skill4, skill6, skill8*
SAH	*skill1, skill2, skill8*
TEC	*skill3, skill5*

4.2. Difference of the skills

In the second part, individual skills have been evaluated and compared. For this analysis, male and female students are not separated into two groups. From the eight skills to be analyzed, two skills (*skill1* and *skill4*) are characterized by the highest average level. Both *skill1* and *skill4* are significantly different from the remaining six skills, while not being significantly different each other. On the other hand, the students have reached the lowest average level for the *skill5*. The mean value is again significantly different from all the other analyzed skills. Table 4 shows order of the skills depending on the average skill level. When two or more skills are not significantly different, they are presented on the same line. As it can be seen, the difference between *skill1* and *skill4*, and *skill2* is only 2%. Due to the fact, that the number of items is high (N = 7 906), this difference is evaluated by the statistic test as significantly different.

Table 4. Average skill levels computed for the skills analyzed in the research.

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

There are only slight differences in the order of the individual skills when the type of school or the sex is considered as an attribute. As we can expect, the values of average level vary for different types of school engaged in the research. This effect is analyzed in the next section.

4.3. Effect of the secondary school

To provide complete analysis of the data, the effect of the school type on the skills has been also evaluated using the presented tool. The average level of the grammar school (GRA) students (both male and female students) is the highest for all the analyzed skills. It is significantly different compared to the other schools. Then, it could be said, that technical schools (TEC) and lyceums (LYC) are characterized with the second highest average level for most of the skills. The values are again significantly different from the remaining schools. The order of the other types of school depends on the concrete skill and no general rule can be derived from the data. Figure 2 shows the Hasse diagram prepared from the data characterizing *skill1*. Grammar schools (GRA) are placed alone on the top of the diagram, which represents the highest average level obtained for the skill. Lyceums (LYC) and technical schools (TEC) are placed together on the same level just below the grammar schools (GRA). Absence of a path between them corresponds to the fact, that there is no significant difference between them for *skill1*.

For *skill2* and *skill7*, the average level obtained for lyceum (LYC) students is significantly different (higher) compared to the average value obtained for technical school (TEC) students.

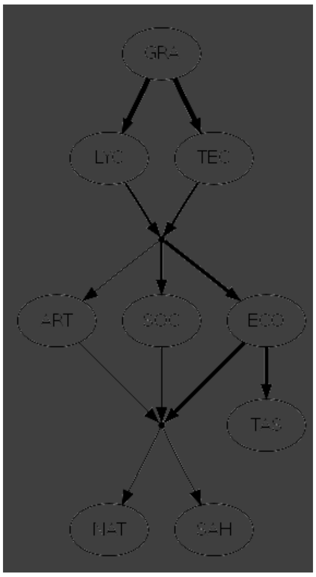

Figure 2. Hasse diagram created for *skill1* (understanding of the number as a concept expressing the quantity)

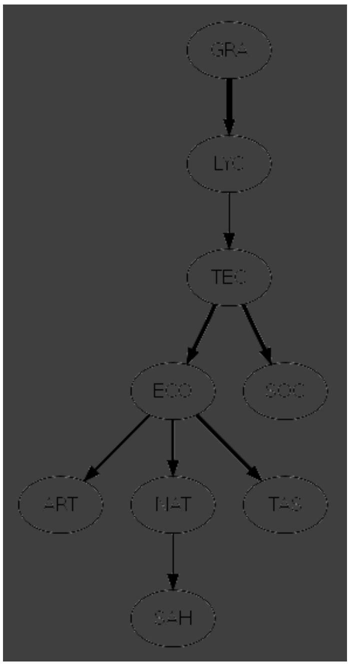

Figure 3. Hasse diagram created for *skill7* (function as relation between quantities)

Only in the case of *skill5*, the result is markedly different. Figure 4 shows the Hasse diagram obtained.

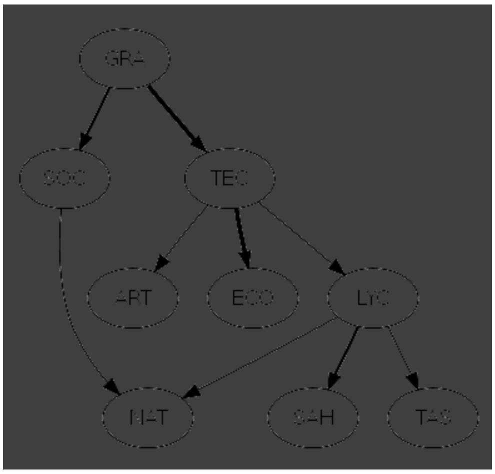

Figure 4. Hasse diagram created for *skill5* (graphical reception and work with graph)

In the next step of the analysis, we focused on evaluation of absolute differences between various types of schools. This analysis shows another two interesting facts. In the case of *skill5*, the difference between the highest average level (grammar school (GRA)) and the lowest average level is only about 8.5%. It represents the smallest difference among the analyzed skills. For grammar schools (GRA), the average skill level reached 46.5%. On the contrary, the worst average level has been obtained for art (ART) and natural science (NAT) and social and health studies (SAH) students (about 38%). This fact strongly corresponds to the results presented in the previous parts, where the average level representing the *skill5* has been determined as very poor compared to the other skills and also the Hasse diagram (Figure 4) representing order of the schools is slightly different.

The greatest difference (over 21%) has been reached for *skill3* and *skill7*. For both the skills, the maximal average level characterizes grammar schools (65% and 64% respectively) and the minimal average level reached art schools (about 43%). In the case of *skill3*, the average level reached for art school is significantly different from the values obtained for other types of school. For the other skills, the difference varies between 14% and 17%.

The variety of absolute difference between types of school is also evident from the diagrams obtained. When the absolute difference is minimal (*skill5*, Figure 4), the structure of the diagram is much wider compared to the skills characterized with maximal absolute difference (e.g. *skill7*, Figure 3). The *skill7* is represented with very narrow structure of the diagram representing the significant differences among averages of the skill levels.

5 Conclusion

We have introduced a new tool for visualization of statistically significant differences among the mean values of quantitative attributes. The method is based on statistical tests of hypotheses of equal means. Firstly, a set of tests is performed in order to determine significant differences among all combinations of tested mean values. The results are then visualized in the Hasse diagram which represents the extracted information in easily understandable format. The proposed technique has been applied on data characterizing mathematical skills of secondary school students. From the results obtained, we can pick up very poor work with graphs (*skill5*) typical for all types of secondary schools.

In the future, the authors of this paper plan to utilize Hasse diagrams to visualize other types of knowledge (e.g. impact rules).

References

[1] Agrawal, R. Fast discovery of association rules. In: Advances in knowledge discovery and data mining. AAAI Press/MIT Press, 1996, 307-328.

[2] Burda, M. Visualization of cosymmetric association rules using Hasse diagrams and concept lattices. In: Znalosti, Hradec Králové, Czech Republic, 2006, ISBN 80-248-1001-8.

[3] Eick, S.G. Visualizing multi-dimensional data. SIGGRAPH Comput. Graph. **34**(1), 2000, 61-67.

[4] Fayyad, U., Piatetsky-Shapiro, G., Smyth, P., Uthursamy, R., eds. Advances in Knowledge Discovery and Data Mining. AAAI Press/MIT Press, 1996.

[5] Fox, J., Friendly, M., Monette, G. Visualizing hypothesis tests in multivariate linear models: the heplots package for R. In: Directions in Statistical Computing, Springer-Verlag, 2008.

[6] Hetzler, B., Harris, W., Havre, S. Visualizing the Full Spectrum of Document Relationships. 1998. [online] http://citeseer.ist.psu.edu/ hetzler98visualizing.html

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

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

[9] NIST/SEMATECH. E-handbook of statistical methods. [online] http://www.itl.nist.gov/div898/handbook/index.htm.

[10] R Development Core Team. R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria, 2008. [online] http://www.r-project.org.

[11] Wong, P.C., Whitney, P., Thomas, J. Visualizing Association Rules for Text Mining. In: INFOVIS, 1999, 120-123.