DE GRUYTER OPEN
Foundations of Management, Vol. 6, No. 3 (2014), ISSN 2080-7279
DOI: 10.1515/fman-2015-0017

DATA MINING AND STATISTICS METHODS FOR ADVANCED TRAINING COURSE QUALITY MEASUREMENT: CASE STUDY

Maxim GALCHENKO*, Alexander GUSHCHINSKY**
Saint-Petersburg State Agrarian University, Saint-Petersburg, Russia
* e-mail: maxim.galchenko@gmail.com
** e-mail: dekan_elfak@mail.ru

Waldemar IZDEBSKI
Warsaw University of Technology, Faculty of Management, Warsaw, Poland
e-mail: w.izdebski@wz.pw.edu.pl

Jacek SKUDLARSKI
Warsaw University of Life Sciences, Warsaw, Poland
e-mail: j.skudlarski@oikos.net.pl

Abstract: Advanced training courses in the energetics field is a very important part of human reliability growth. In the words of S.E. Magid, chief of Technical Educational systems in Energy Technologies, UNESCO department: "The number of forced outages due to failures of equipment on the power stations -30 %. The share of operational personnel fault in these infringements makes considerable size (to 15%). As a whole in the Russian Open Society 'United Power Systems' the infringements percentage because of the personnel from infringements total makes 2%. At the same time, on power stations this quantity makes 18%. In power supply systems of Siberia the relative quantity of infringements because of the personnel reaches 50%." [1].

Keywords: advanced training course, educational data mining, statistics methods.

1 Data and methods

In this case, the HR department must receive students' response and adequately analyze it. For this purpose, the HR department of one of the largest power providers prepared a questionnaire for students (electrician, electric engineers) who have a course under a specially organized Advanced Training Courses department. In the questionnaire, students marked some parameters of courses, which were re-coded for analysis purposes (Table 1). Every parameter was marked in the diapason from 1 (bad) to 10 (excellent). Questionnaires were deanonymized (we did not have access to the questionnaire preparation process, so we cannot adequately explain this fact).

It should be noted that the last parameter "General impressions about provided training" in fact is generalizing for all others (a target variable), that is, in Educational Data Mining (EDM) terminology – label.

The main goals of analysis were formulated by the HR department as:

- to estimate course quality,
- to identify the major factors influencing an assessment of a course by students,
- to make recommendations about improvement of course characteristics, the questionnaire improvements.

For the analysis we received 316 questionnaires completed by students at the end of the advanced training course. Initially data underwent cleaning for incomplete cases, because the applied methods could not process them. As a result, the dataset decreased to 301 records.

We carried out processing only in open source software: statistical programming language R [2] and data mining platform of KNIME [3].

2 Data analysis and results

Verification of normality in all fields across the dataset by the Shapiro–Wilk test gives a negative result at the 0.05 p-level. The peak is detected at level 10 (excellent) for all fields. So, for an indicator "General impressions about provided training," the histogram shows value 10 prevailing, and the values lower than 6 frequencies (see Fig. 1) are very rare, and that is characteristic for all other fields.

Table 1. Parameters and their coding for the processing

Parameter	Coding
General organization of training process	obshorg
Equipment of educational audience	oboryd
Compliance time of courses with syllabus	cootvraspis
Compliance of the course content with the declared program	sootvprogr
Practical usefulness of the carried-out training	polezn
Comprehension of a course material	dostypmat
Use of modern methods in a process	sovrmetod
Individual approach to the student	indpodxod
Efficiency of interaction of the teacher with audience	vzaimodayd
Receiving feedback from teachers	obrcvyazprepod
Receiving feedback from employees of chair	obrcvyazkaf
Interaction of students among themselves	vzaimodslysh
Existence, quality, usefulness of printing materials	razdatmat
Quality of food in the dining room	pitanie
General impressions about provided training	obsh

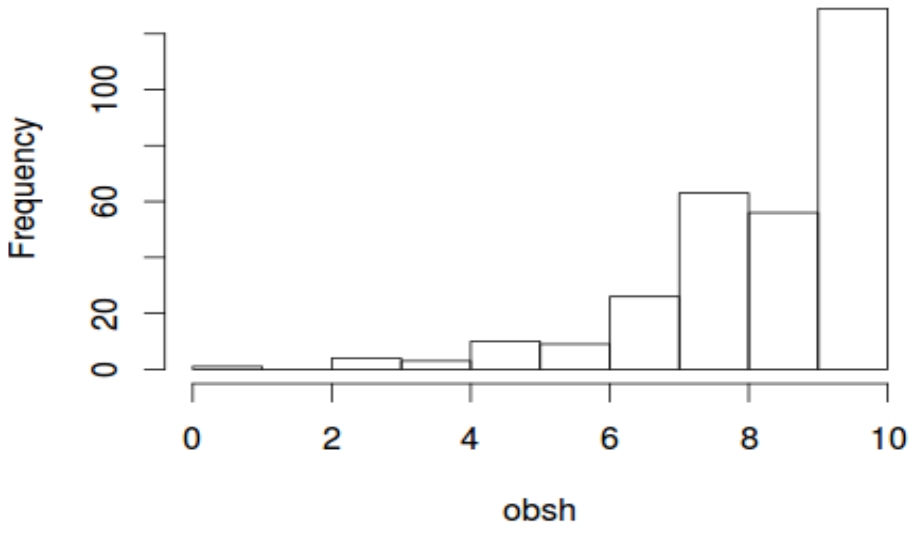

Figure 1. The histogram of the field "General impressions about provided training"

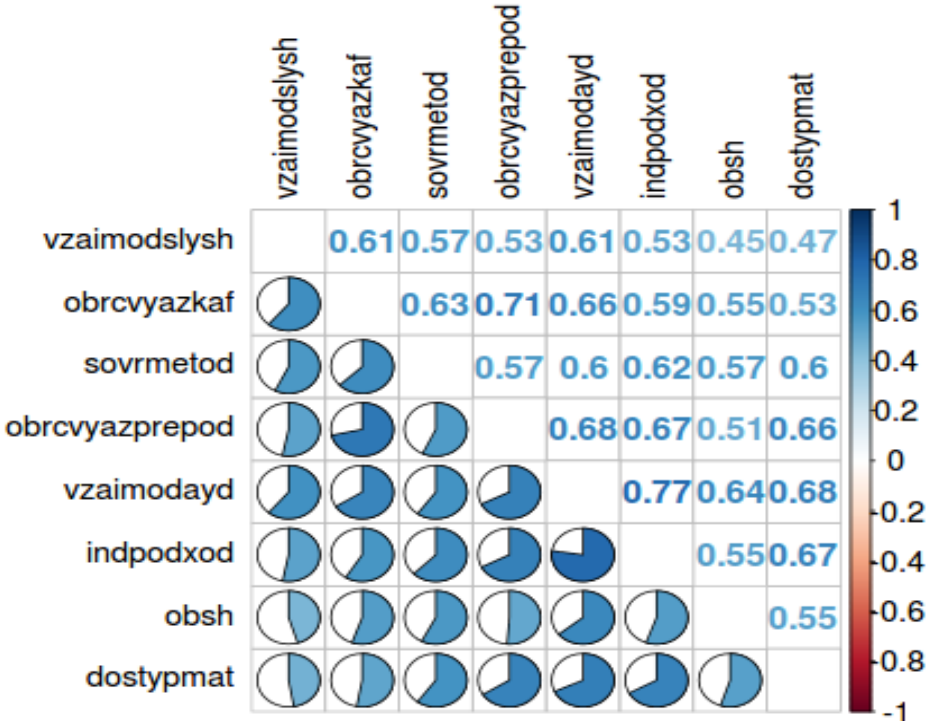

Figure 2. Correlation matrix

So, we can use only non-parametric statistics methods.

The data is ordinal, so we used Spearmen correlations subsequently subjected to a filtration at the level of 0.6 with corrplot package [4] of the R language (see Fig. 2).

All correlation coefficients are significant at the 0.05 p-level.

From the results it is possible to claim that there is a high level of correlation between factors "An individual approach to the student" and "Efficiency of interaction of the teacher with audience," and "Receiving feedback from teachers" and "Receiving feedback from employees of chair." Thus, it is possible to say with confidence that students practically do not distinguish between an individual approach and efficiency of the teacher, and marks feedback from teacher and feedback from chair concordantly. It is really interesting that “Individual approach to the student” and “Comprehension of a course material” are highly correlated.

It should be noted a large share of correlations in the dataset with values that can be treated as “average” (values from 0.5 to 0.7).

In the scoring process can be the groups of persons formed, operating in a similar way; therefore obtaining information on the existence or absence of such groups was the following issue, which was resolved within an objective.

The procedure of the cluster analysis was applied for these purposes to the dataset with fuzzy C-means with the predetermined quantity of clusters equal to 4. Four groups, one of which came to about 40%, were as a result received, and other groups comprised 17% to 20% of the power of the initial dataset (Table 2).

For distinction analysis among the received clusters, medians and interquartile range (IQR) on all indicators (Table 3) were calculated.

Table 2. Respondents’ answer distribution by clusters

Cluster	Elements	Percent
1	58	19%
2	117	39%
3	75	25%
4	51	17%
Total	301	100%

Table 3. IQR and medians for all parameters by clusters

Parameter/cluster	IQR, by clusters				Median, by clusters			
	1	2	3	4	1	2	3	4
General organization of training process	1.0	0.0	2.0	1.0	7.5	10.0	9.0	8.0
Equipment of educational audience	3.0	1.0	3.5	2.0	7.0	10.0	8.0	8.0
Compliance time of courses with syllabus	2.0	0.0	0.5	1.0	9.0	10.0	10.0	10.0
Compliance of the course content with the declared program	1.0	0.0	1.0	2.0	8.0	10.0	10.0	9.0
Practical usefulness of the carried-out training	3.0	0.0	2.5	2.0	7.0	10.0	8.0	8.0
Comprehension of a course material	1.0	0.0	1.0	1.0	8.0	10.0	10.0	8.0
Use of modern methods in a process	1.8	0.0	1.0	1.0	7.0	10.0	10.0	8.0
Individual approach to the student	2.0	0.0	1.0	2.0	7.0	10.0	10.0	8.0
Efficiency of interaction of the teacher with audience	1.0	0.0	1.0	1.0	8.0	10.0	10.0	9.0
Receiving feedback from teachers	1.0	0.0	0.0	2.0	8.0	10.0	10.0	9.0
Receiving feedback from employees of chair	2.0	0.0	0.0	2.0	7.0	10.0	10.0	9.0
Interaction of students among themselves	1.0	0.0	0.0	2.0	8.0	10.0	10.0	9.0
Existence, quality, usefulness of printing materials	3.0	0.0	3.0	2.0	6.5	10.0	8.0	8.0
Quality of food in the dining room	2.8	1.0	2.5	4.0	7. 0	10.0	9.0	8.0
General impressions about provided training	2.0	0.0	2.0	1.5	7.0	10.0	9.0	8.0

As a whole, it should be noted high marks for all parameters by students of clusters 2–4. In fact, we must remember that questionnaires were deanonymized, so it make sense in this case and we can expect, that marks will be higher than in anonymized questionnaires.

Here we can draw some conclusions that the available information allows us to make.

The cluster 2 students mark 13 parameters as “excellent” and "Equipment and the equipment of educational audience" and "Quality of food in the dining room" in most cases on 9–10 points. Most likely, this group of students approached formally the questionnaire filling process and has to be excluded from further analysis. We think that, this cluster was formed by “panic students” who did not add the real marks in the questionnaire with their names at the top.

Cluster 1 consists of most judicial students. Especially, it should be noted a low mark in the "Existence, quality, usefulness of printing materials" parameter in this group. In general, with respect to printing materials, distinctions between groups on this factor are significant at the 0.05 p-level (Kruskal–Wallis chi-squared = 20.5337, df = 2, p = 3.477e−05). More clearly it can be shown on a boxplot (see Fig. 3).

A similar pattern is observed for the "Quality of food in the dining room" parameter. Distinctions between groups on this factor are significant at the level of 0.05 (Kruskal-Wallis chi-squared = 20.0733, df = 2, p = 4.377e−05); thus in cluster 4, IQR is very large, the first quartile is equal to 5 points (see Fig. 4).

Distinctions on a parameter "Practical usefulness of the carried-out training" are also statistically significant at the level of 0.05 (Kruskal–Wallis chi-squared = 38.5622, df = 2, p = 4.23e−09); thus 25% of the first group marked usefulness lower than 5 (see Fig. 5).

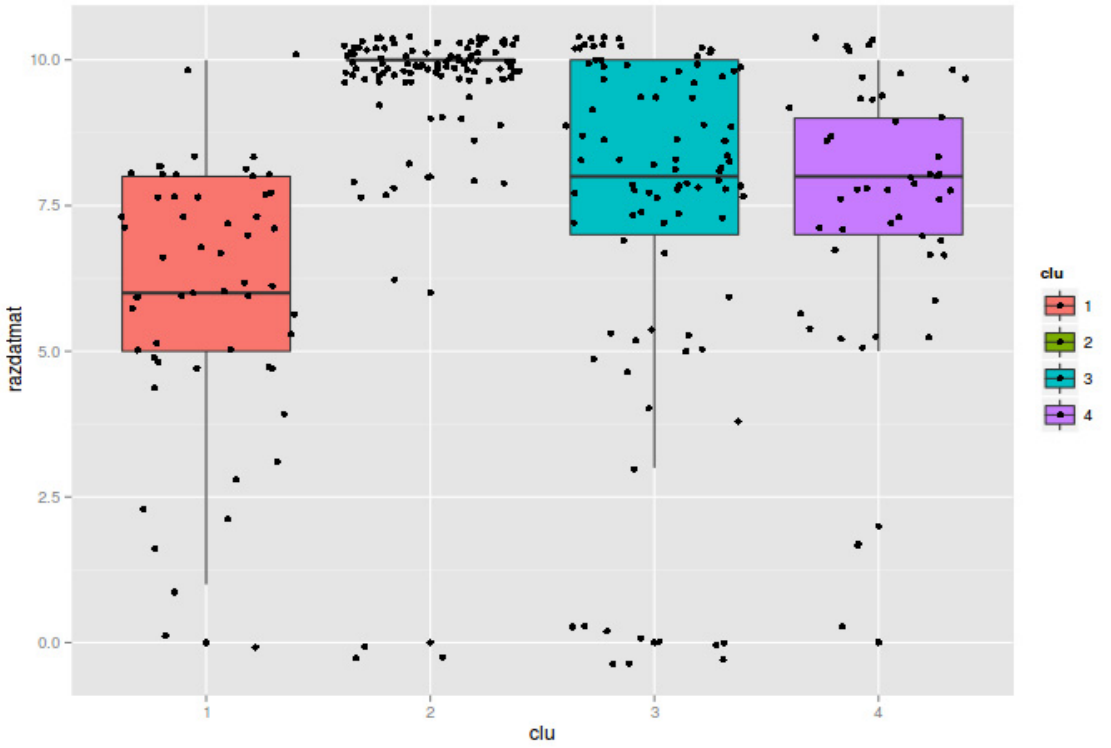

Figure 3. Printing material scoring in clusters 1–4

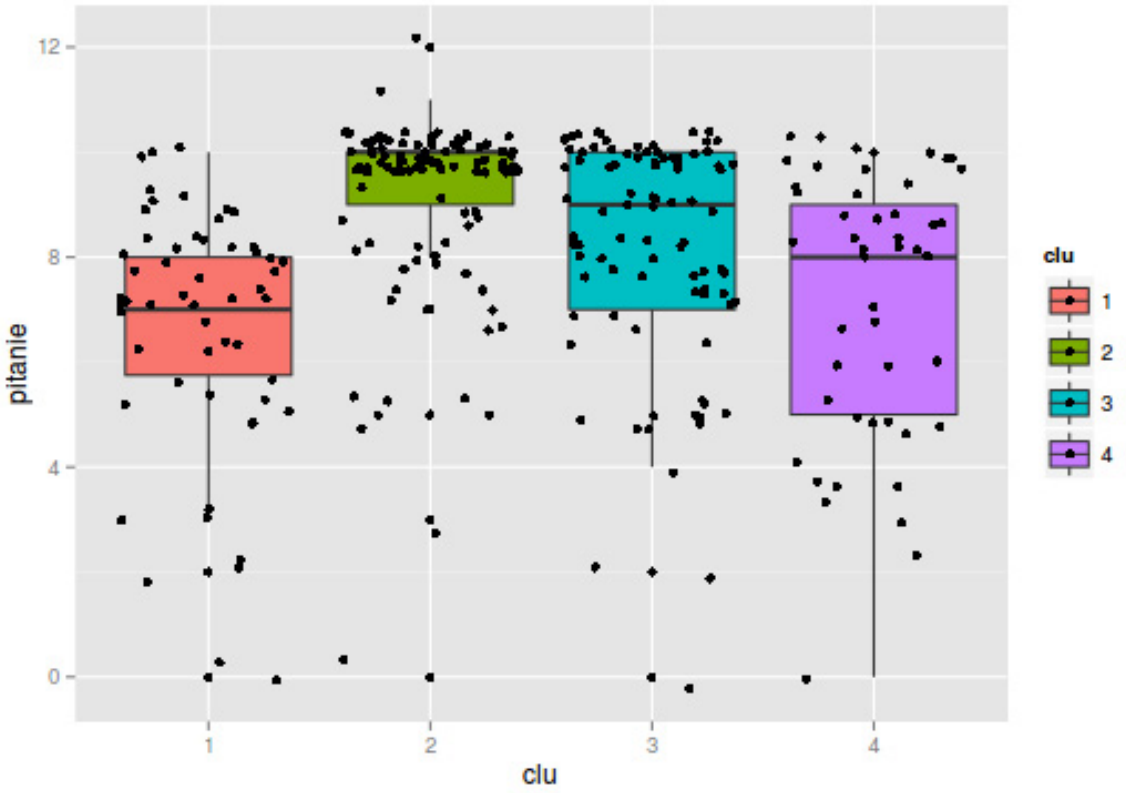

Figure 4. Marks for "Quality of food in the dining room" in various groups

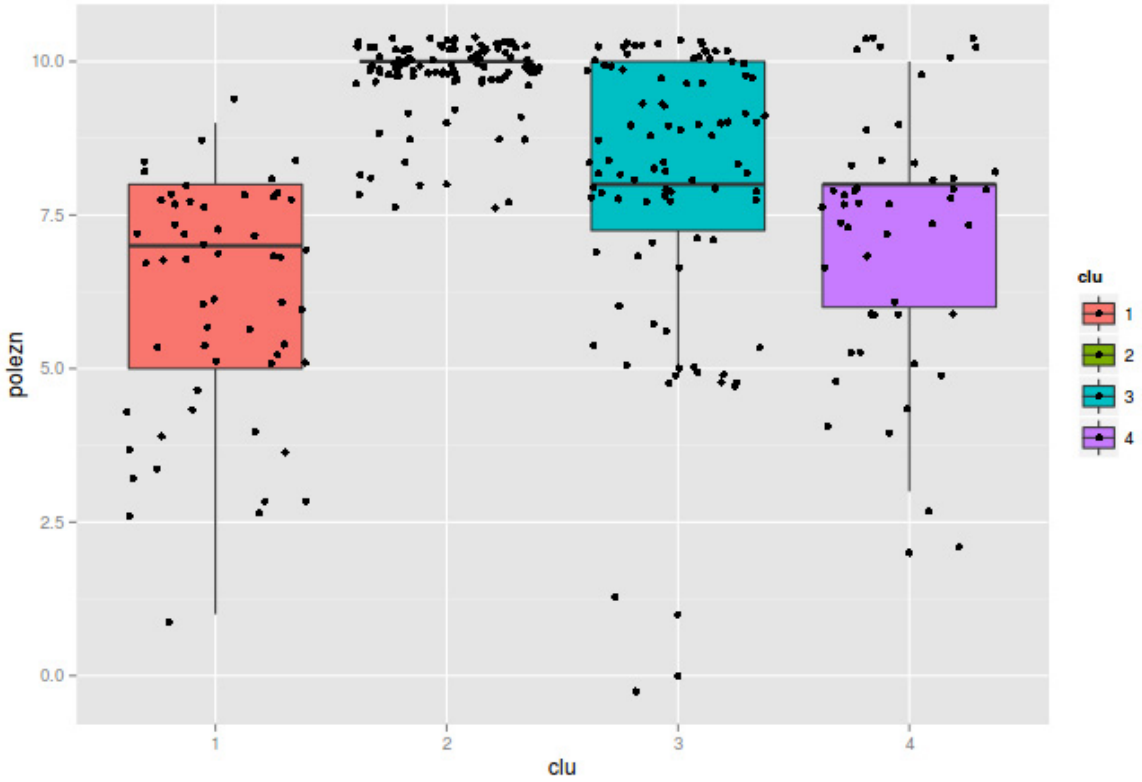

Figure 5. Marks for "Practical usefulness of the carried-out training" in various groups

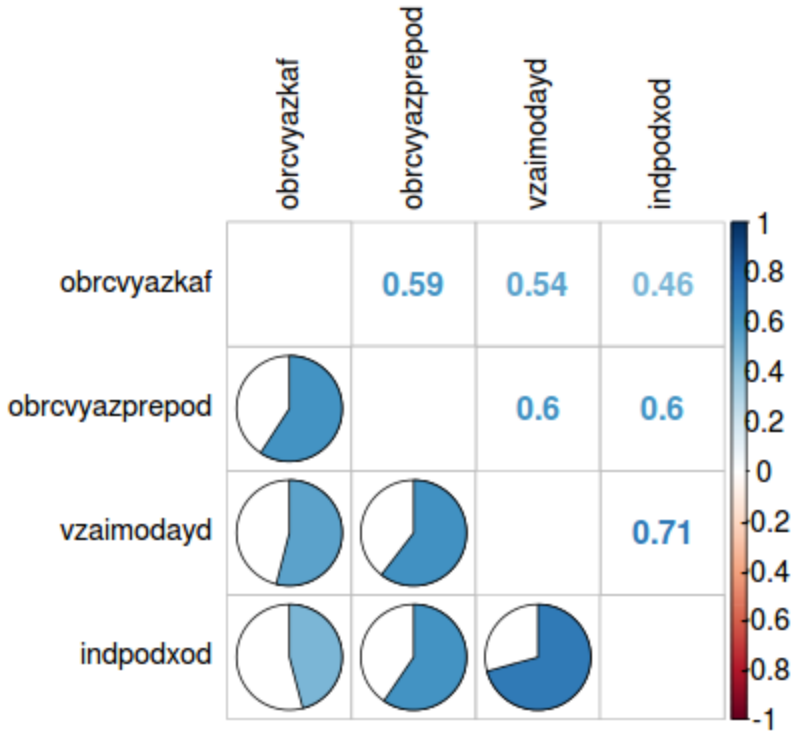

Figure 6. Correlation matrix with 0.6 threshold after a filtration of cluster 2

Table 4. Factor loadings (in variation explained order)

Parameters	MR1 (Factor 1)	MR2 (Factor 2)	MR4 (Factor 4)	MR3 (Factor 3)	MR5 (Factor 5)
General organization of training process	−0.003	0.717	0.289	0.228	0.026
Equipment of educational audience	0.137	0.501	−0.028	0.002	0.021
Compliance time of courses with syllabus	0.001	0.416	0.175	0.035	0.358
Compliance of the course content with the declared program	0.332	0.131	0.554	−0.011	0.327
Practical usefulness of the carried-out training	0.228	0.135	0.468	0.050	0.066
Comprehension of a course material	0.571	0.135	0.501	−0.009	−0.048
Use of modern methods in a process	0.203	0.693	0.274	−0.135	−0.063
Individual approach to the student	0.786	0.187	0.143	−0.066	−0.029
Efficiency of interaction of the teacher with audience	0.761	0.169	0.166	0.147	0.241
Receiving feedback from teachers	0.654	0.192	0.361	0.147	0.136
Receiving feedback from employees of chair	0.254	0.596	0.063	0.194	0.273
Interaction of students among themselves	0.186	0.547	0.042	−0.009	0.301
Existence, quality, usefulness of printing materials	0.068	0.079	0.074	−0.221	0.326
Quality of food in a dining room	0.101	0.109	0.059	0.716	−0.096

A similar pattern is observed for "Equipment of educational audience," "Comprehension of a course material," and "Individual approach to the student" parameters.

About a third of respondents are in clusters 1 and 4, in which lower marks than those described above are observed for most of the parameters.

For identification of the hidden factors and their relative importance, the factor analysis was used. Previously cluster 2 was excluded from analysis, as mensionbefore.

The filtration led to the loss of a significant part of linear correlations (see Fig. 6).

The analysis of statistically significant parameters leads to the conclusion that students closely coordinate "An individual approach to the student," "Efficiency of interaction of the teacher with audience," and "Receiving feedback from teachers" (correlation coefficient not less than 0.6). The average level of communication (0.59) between factors "Receiving feedback from teachers" and "Receiving feedback from employees of chair" remains.

Results of the factor analysis (psych package [5]) allow us to assume the existence of five hidden factors (Maximum Likelihood chi-squared = 39.804, p <t; 0.134) (see Table 4).

These five factors explain about 50% of a dataset variation; thus factors 1 and 2 explain 16% of a variation, and factors 4, 3 and 5explain about 8%, 5%, and 4% of variation respectively.

The graphic interpretation of results (see Fig. 7) gives more clearly the structure factor composition.

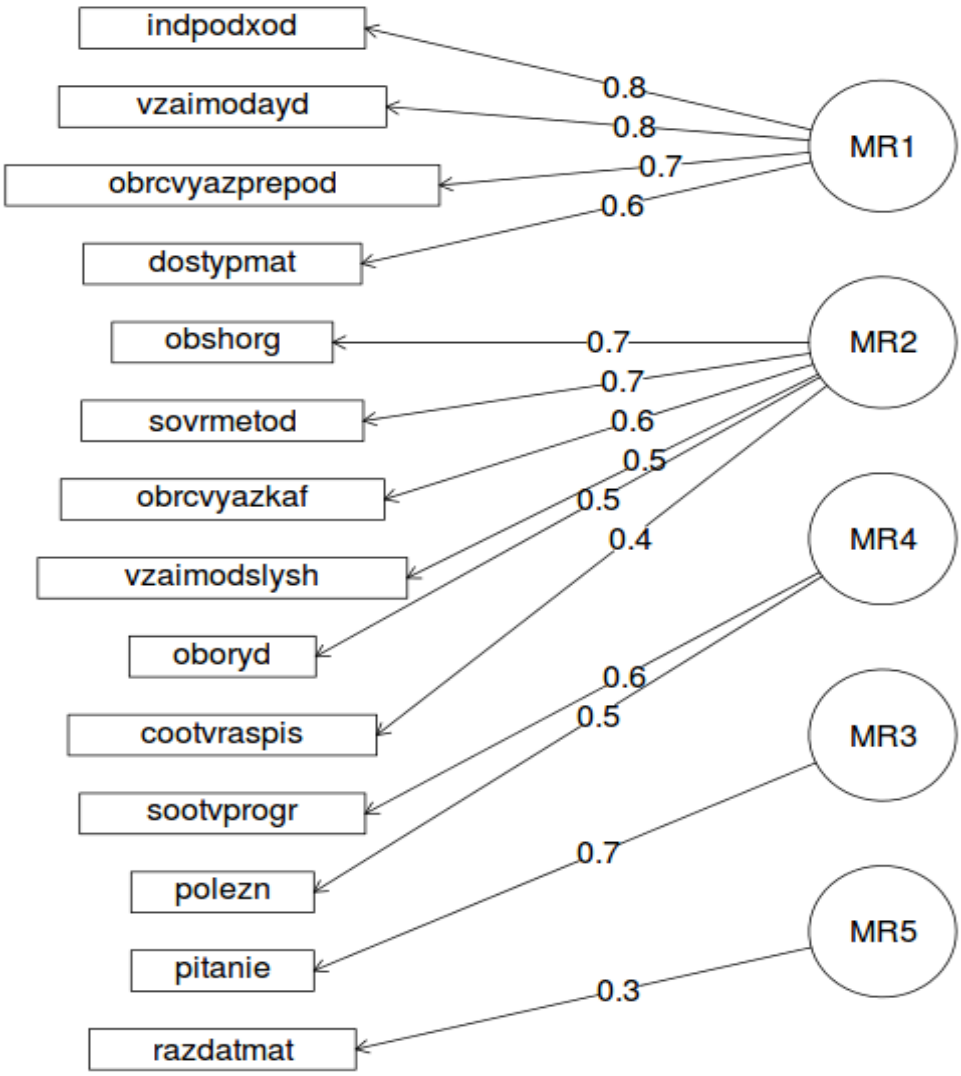

Figure 7. Factor structure

The first factor can be treated as "Individual skill of the teacher" (individual skill, effective interaction with students, comprehension of a course material), the second – "Overall performance of chair" (the general organization, modern methods of teaching, feedback, interaction of students, the equipment, compliance with the syllabus), and the third – "Usefulness of training" (compliance with the program, usefulness). Separately there is a quality of food and printing materials.

According to the ranging of factors, individual skill of the teacher and overall performance of chair have the greatest weight. The management aimed at the solution of these problems will have the greatest impact due to the result of factor analysis. Usefulness is also a sufficiently important factor, which demands special attention, taking into account the results received earlier.

We research the possibility of forecasting the “General impressions about provided training” parameter, trying various algorithms of classification in the KNIME data mining platform.

A scale change for the field "General impressions about provided training" for the purpose of increase in frequencies by possible versions of the answer was carried out:

- values from 0 to 6 were coded by 1 (Bad),
- from 7 to 8 – 2 (Well),
- from 9 to 10 – 3 (Perfect).

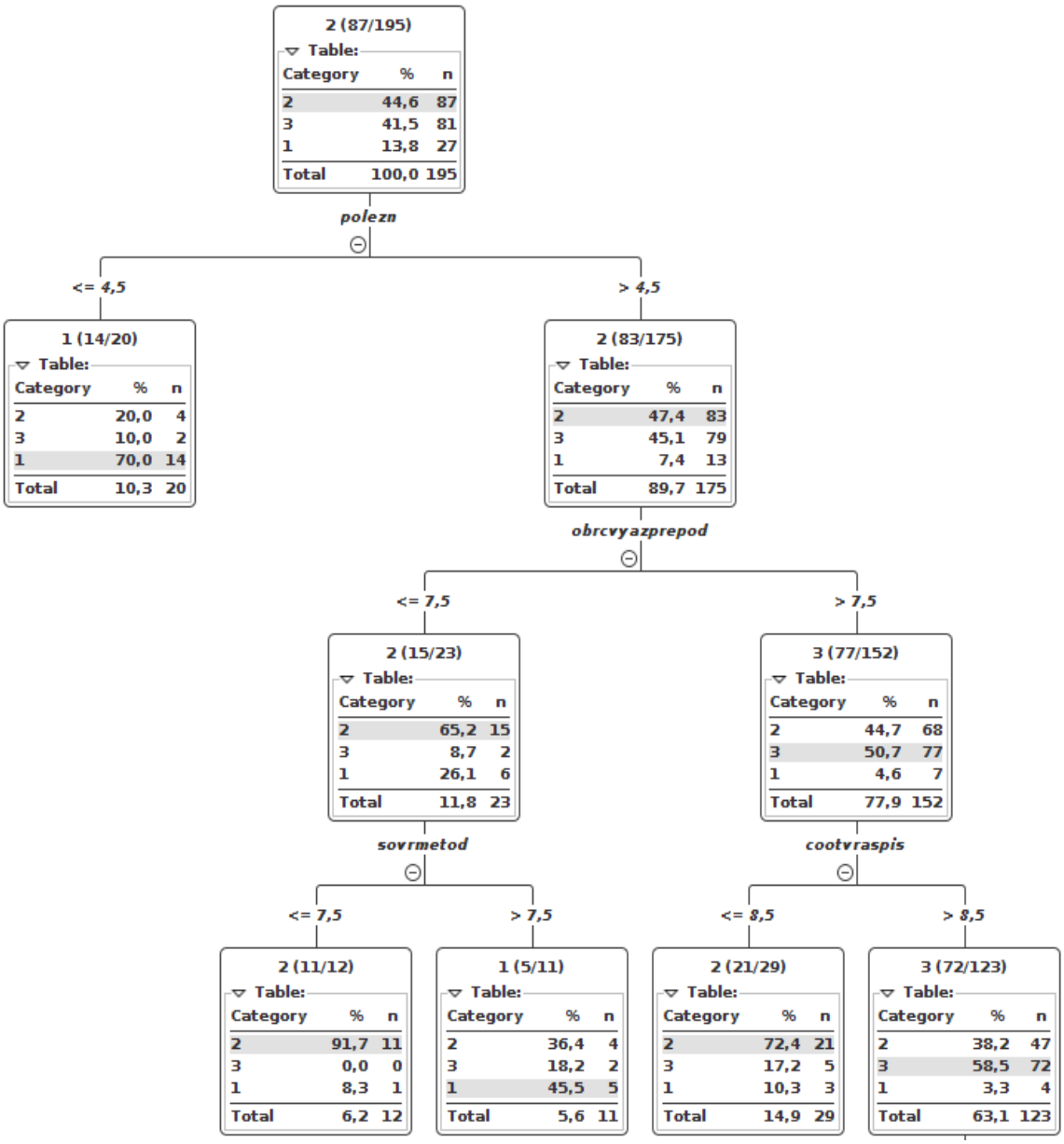

Figure 8. Decision tree, using "General impressions about provided training" as target variable

The result is not really impressive: the best accuracy in cross-validation by the leave-one-out method is 63% given by the decision tree algorithm (a quality measure – gain ratio, the minimum quantity of records per node is 10). The value of F-measure on 20% test dataset yielded a rather good result for class "1" (0.8) and satisfactory for classes "2" and "3" (0.6 and 0.632, respectively, which is explained by low specificity of class "3" – 0.565 and low sensitivity of class "2" – 0.529). The resulting model provides structure of course scoring by participants in general (see Fig. 8).

So (Fig. 8), about a half of students (17 of 24), scored from 0 to 6 “General impressions from a course” marked a course useless for themselves (marks lower than 5 points for “Practical usefulness of the carried-out training”).

On the other hand, the usefulness was estimated at 9–10 points, feedback with chair at 9–10 points, compliance of occupations with the schedule at 9–10 points, with the teacher at 8–10 points giving to feedback in the general impression at 9–10 points (84%, 36 of 43 students).

If, when performing all previous conditions, usefulness is scored in the range from 6 to 8 points, the main role is played by food quality and the general organization of the process: the food quality scored at 8–10 points and the general organization, estimated at 9–10 points yields the result "Perfect" (73%, 19 of 26 students). If food quality is scored at 7 points and below, the result in most cases is "satisfactory" (78%, 14 of 18 students).

Other ranges of the factors stated above yield an unstable result in view of small frequencies.

3 Summary

On the basis of the analysis carried out, it is possible to formulate the following recommendations:

- To focus attention of students on thoughtful filling of the questionnaire as a large number of questionnaires go to waste and the final reduces to 40% from the original number.
- Electronic form of testing will be applied and new fields must be added (age, experience, education, and so on), because quality of the prediction model is low. Questionnaire anonymization is needed.
- Prime parameters for improvement can be allocated: "Existence, quality, usefulness of printing materials," "Quality of food in the dining room," "Practical usefulness of the carried-out training," "Equipment of educational audience," "Comprehension of a course material," and "Individual approach to the student".
- When hiring teachers, it is worth paying attention to readiness to work in a schedule grid, the maximum aim at individual work, effective interaction with students.
- Food, the general organization of process, and effective feedback with chair are capable of increasing the general satisfaction of students significantly.
- Especially, it should be noted the importance of factors: "Quality of food in the dining room" and "General organization of training process" when the usefulness of direct courses is not obvious to students.

4 References

[1] Магид С.И. - Человеческий фактор и энергобезопасность на современном этапе реформирования электроэнергетики РФ. [online] [10.12. 2013.]. Available at: http://www.testenergo.ru/003.rtf

[2] R CORE TEAM (2013) R: A language and environment for statistical computing. R Foundation for Statistical Computing//Vienna, Austria. URL http://www.R-project.org/.

[3] Michael R. Berthold, Nicolas Cebron, Fabian Dill, Thomas R. Gabriel, Tobias Kotter, Thorsten Meinl, Peter Ohl, Christoph Sieb, Kilian Thiel, Bernd Wiswedel - Studies in classification, data analysis, and knowledge organization. knime: the konstanz information miner//Springer, 2007.

[4] Taiyun Wei corrplot: Visualization of a correlation matrix. R package version 0.73.//http://CRAN.R-project.org/package=corrplot.

[5] Revelle W. psych: Procedures for Personality and Psychological Research//Northwestern.

[6] University, Evanston, Illinois, USA, http://cran.r-project.org/web/packages/psych/index.html.