USING CLUSTER ANALYSIS FOR THE STUDY OF GENERATION Y BEHAVIOR AT THE UNIVERSITY OF PARDUBICE

Hana Jonášová, Karel Michálek, Jan Panuš

Abstract: *For Generation Y are social networks an essential element of communication and entertainment. The authors of this article conducted research in the Faculty of Economics and Administration at the University of Pardubice on how students (generation Y) use social networks. For the evaluation of research were used cluster analysis and decision trees.*

Keywords: *Generation Y, Social network, Faculty of Economics and Administration University of Pardubice, Facebook, Communication, Questionnaire survey, Cluster analysis, Kohonen self-organizing feature maps, K-Means, Two-Step, C5.0, Decision trees.*

JEL Classification: *C15, C44, C63, Y10.*

Introduction

Generation Y [2], [5], [10] is shaped by the period in which they live. The generation grows up surrounded by modern technology that is, of course, completely used from an early age. The global expansion of the Internet and mobile networks was very import for the formation of this generation.

We were interested in the specifics Generation Y of the Faculty of Economics and Administration, University of Pardubice. Therefore we made a questionnaire survey to clarify how this generation communicates with each other and how uses the means social networks. We focused mainly on selected internet social networks in this study. The contribution of this article is within describing of this generation and the network by statistical methods and methods of cluster analysis.

1 Problem formulation

This paper builds on an article [4] that focused on the preparation of questionnaires for this research and primary processing. We discussed the status of social networks and their influence on Generation Y in the University of Pardubice in this article. This article raised many more questions for us. How to describe the structure of social networks? What is the relationship between the social networks? Who are the typical users at the University of Pardubice?

We made two rounds of questionnaire survey as described in the [4]. We asked respondents how often they use social networks for their communication and how much time they spend with the computer in the first round. Respondents were asked to describe their own social network in the second round. We asked 182 respondents from Faculty of Economics and Administration, University of Pardubice.

Questions were classified to two subscales a) importance of the internet, b) using web services as communication or for leisure activities.

2 Problem solving

The questionnaire process, primary data pre-processing and rough summary was published in [4]. We decided for additional data processing to use cluster analysis. This enables evaluating and understanding data obtained from the survey. The method for understanding the various clusters was decision trees. The combination of these methods is often used. Using the combination of these techniques in relation to social networks is described in [1].

2.1 Data preprocessing

Data must be transformed from the individual questionnaires to a CSV file for processing in software Clementine 10.1. These data were adjusted manually and then using the scaling function of MS Excel covered in the appropriate form.

The next step for data preprocessing and analysis was necessary to treat some of the characteristics. These were mainly about Internet services, which were represented in the questionnaire with near zero frequency.

We removed the following services: Lidé.cz; Líbímseti.cz; LinkedIn; Twitter; Orkut; Flickr; MySpace; N-JOY.cz; Vimeo; Picasa; Rajče.net; Facebook games.

Cluster Analysis

Therefore, to understand semantics data obtained from the research of the Kohonen networks were used in order to determine the number of primary clusters. Cluster analysis was performed in software Clementine 10.1. It was based on data text file (CSV) which was processed in MS Excel. We used Two-Step algorithm and K-Means method for additional processing.

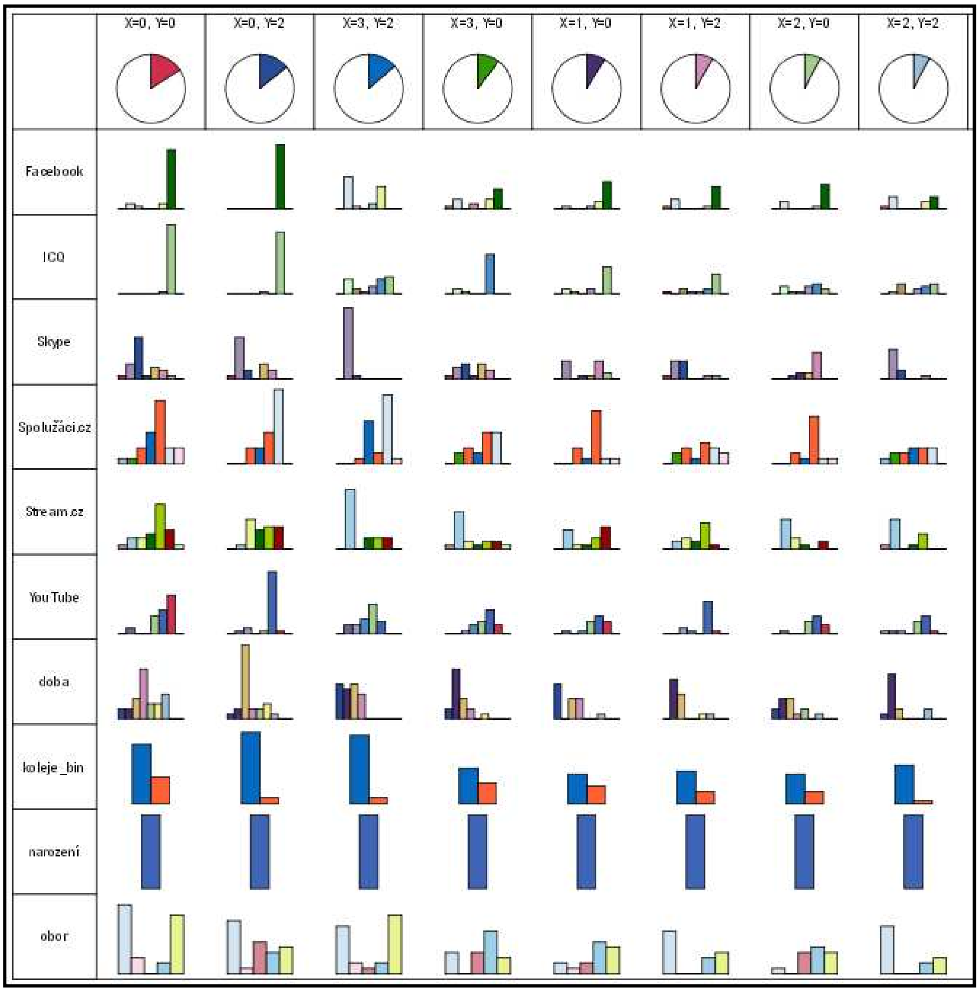
2.2 Kohonen self-organizing feature maps

Kohonen self-organizing feature maps [6] are a type of neural network based on competitive learning strategy, the input layer serves the distribution of the input patterns. The neurons in the competitive layer (output layer) serve as representatives, and they are organized into topological structure. All the input neurons are connected to all of the output neurons, and between these connections are weights, associated with them and there is distances computed between them. The winning neuron is chosen, for which the distance from input pattern is minimum.

This type of network can be used for clustering data sets into distinct groups when unsupervised method is used. Records are grouped so that records within a group or cluster tend to be similar to each other, and records in different groups are dissimilar.

The result of cluster analysis using Kohonen maps is given in Fig. 1, it describes the different clusters and the various input variables for different clusters (Facebook, ICQ, Skype, Spolužáci.cz, You Tube, time spent at the computer, college dormitory, year of birth, field of study).

Fig. 1: The result of cluster analysis using Kohonen maps



Source: own

Numbers of elements in clusters, created by Kohonen maps are given in Tab. 1 Cluster analysis using Kohonen maps created 12 clusters. But the total numbers of 12 clusters are only 3 clusters significant (highlighted cells). The number of clusters will be further used for other clustering methods.

Tab. 1: Number of clusters

Y, X	0	1	2	3
0	29	16	14	19
1	5	4	6	9
2	26	14	14	25

Source: own

2.3 Two-Step Algorithm

Two-Step Cluster is a two-way clustering method. The first step makes a single pass through the data, during which it compresses the raw input data into a manageable set of sub clusters. The second step uses a hierarchical clustering method to progressively merge the sub clusters into larger and larger clusters, without requiring another pass through the data. Hierarchical clustering has the advantage of not requiring the number of clusters to be selected ahead of time. Many hierarchical clustering methods start with individual records as starting clusters and merge them recursively to produce ever larger clusters. Though such approaches often break down with large amounts of data, Two-step's initial pre-clustering makes hierarchical clustering fast even for large data sets.

Two-Step clustering method could be used, since using Kohonen map was the chosen number of clusters. For easier interpretation of cluster analysis 3 target clusters was selected. The results of cluster analysis are the following tables. The tables (Tab. 2 – 7) show percentage of respondents belonging to the cluster, and what value is dominant. The rows of table are scale, how often is the social network used. Highlighted cells are dominant for the cluster. The tables (Tab. 8 – 10) show the results from another perspective. Individual monitored variables are in the rows. Highlighted cells are dominant for the cluster.

Tab. 2: Results for Facebook

Facebook	cluster 1	cluster 2	cluster 3
0	2,35%	0,00%	3,7%
1	30,59%	2,86%	29,63%
2	2,35%	1,43%	0,00%
3	1,18%	1,43%	0,00%
4	4,71%	0,00%	0,00%
5	25,88%	8,57%	3,7%
6	32,94%	85,71%	62,96%

Source: own

Tab. 3: Results for ICQ

ICQ	cluster 1	cluster 2	cluster 3
0	1,18%	0,00%	0,00%
1	16,47%	0,00%	7,41%
2	12,94%	0,00%	3,70%
3	4,71%	2,86%	0,00%
4	10,59%	5,71%	0,00%
5	32,94%	8,57%	18,52%
6	21,18%	82,86%	70,37%

Source: own

Tab. 4: Results for Skype.

Skype	cluster 1	cluster 2	cluster 3
0	3,53%	1,43%	3,7%
1	56,47%	20,00%	66,67%
2	12,94%	27,14%	25,93%
3	3,53%	2,86%	0,00%
4	4,71%	21,43%	0,00%
5	17,65%	21,43%	3,70%
6	1,18%	5,71%	0,00%

Source: own

Tab. 5: Results for Spolužáci.cz.

Spolužáci	cluster 1	cluster 2	cluster 3
0	2,35%	2,86%	0,00%
1	7,06%	1,43%	3,70%
2	18,82%	11,43%	3,70%
3	18,82%	18,57%	0,00%
4	32,94%	42,86%	0,00%
5	20,00%	15,71%	81,48%
6	0,00%	7,14%	11,11%

Source: own

Tab. 6: Results for Stream.cz.

Stream.cz	cluster 1	cluster 2	cluster 3
0	2,35%	1,43%	0,00%
1	58,82%	14,29%	7,41%
2	5,88%	17,14%	22,22%
3	9,41%	7,14%	33,33%
4	20%	31,43%	3,70%
5	3,53%	25,71%	25,93%
6	0,00%	2,86%	7,41%

Source: own

Tab. 7: Results for YouTube.cz..

YouTube	cluster 1	cluster 2	cluster 3
0	1,18%	0,00%	0,00%
1	1,18%	5,71%	0,00%
2	8,24%	1,43%	11,11%
3	12,94%	0,00%	7,41%
4	30,59%	17,14%	0,00%
5	30,59%	40,00%	81,48%
6	7,06%	35,71%	0,00%

Source: own

Tab. 8: Sex of the respondent of cluster analysis.

Sex	cluster 1	cluster 2	cluster 3
men	25,88%	28,57%	3,70%
woman	74,12%	71,43%	96,3%

Source: own

Tab. 9: Grade of cluster analysis.

Grade	cluster 1	cluster 2	cluster 3
1	0%	2,86%	3,70%
2	96,47%	87,14%	88,89%
3	3,53%	10%	7,41%

Source: own

Tab. 10: Time spent at the computer of cluster analysis.

Time	cluster 1	cluster 2	cluster 3
1	22,35%	11,43%	0,00%
2	41,18%	10,00%	22,22%
3	18,82%	24,29%	59,26%
4	10,59%	25,71%	0,00%
5	1,18%	10,00%	3,7%
6	1,18%	8,57%	11,11%
7	4,71%	10,00%	3,70%

Source: own

2.4 K-Means method

K-Means [7] defines a set of starting cluster centres derived from data. It then assigns each record to the cluster to which it is most similar, based on the record's input field values. After all cases have been assigned, the cluster centres are updated to reflect the new set of records assigned to each cluster. The records are checked again to see whether they should be reassigned to a different cluster, and the record assignment/cluster iteration process continues until either the maximum number of iterations is reached, or the change between one iteration and the next fails to exceed a specified threshold. The tables (Tab. 11 - 16) show the percentage of respondents belonging to the cluster, and what value is dominant. The rows of table are scale; it shows how often the social network is used. Highlighted cells are dominant for the cluster. The tables (Tab. 17 – 19) show the results from another perspective. Individual monitored variables are in the rows. Highlighted cells are dominant for the cluster.

Tab. 11: Results for Facebook.

Facebook	cluster 1	cluster 2	cluster 3
0	0,00%	2,17%	4,76%
1	10,64%	34,78%	23,81%
2	2,13%	0,00%	2,38%
3	1,06%	0,00%	2,38%
4	2,13%	0,00%	4,76%
5	20,21%	10,87%	11,90%
6	63,83%	52,17%	50,00%

Source: own

Tab. 12: Results for ICQ.

ICQ	cluster 1	cluster 2	cluster 3
0	1,06%	0,00%	0,00%
1	11,70%	2,17%	9,52%
2	8,51%	0,00%	9,52%
3	5,32%	0,00%	2,38%
4	8,51%	2,17%	9,52%
5	28,72%	6,52%	21,43%
6	36,17%	89,13%	47,62%

Source: own

Tab. 13: Results for Skype.

Skype	cluster 1	cluster 2	cluster 3
0	3,19%	2,17%	2,38%
1	38,30%	56,52%	42,86%
2	21,28%	21,74%	16,67%
3	5,32%	0,00%	0,00%
4	10,64%	6,52%	14,29%
5	20,21%	8,70%	19,05%
6	1,06%	4,35%	4,76%

Source: own

Tab. 14: Results for Spolužáci.cz.

Spolužáci	cluster 1	cluster 2	cluster 3
0	1,06%	2,17%	4,76%
1	3,19%	2,17%	4,76%
2	11,70%	17,39%	14,29%
3	17,02%	4,35%	26,19%
4	44,68%	2,17%	35,71%
5	17,02%	65,22%	9,52%
6	5,32%	6,52%	0,00%

Source: own

Tab. 15: Results for Stream.cz.

Stream.cz	cluster 1	cluster 2	cluster 3
0	0,00%	0,00%	7,14%
1	39,36%	30,43%	26,19%
2	8,51%	15,22%	19,05%
3	7,45%	17,39%	16,67%
4	27,66%	8,7%	23,81%
5	15,96%	23,91%	4,76%
6	1,06%	4,35%	2,38%

Source: own

Tab. 16: Results for YouTube.cz.

YouTube	cluster 1	cluster 2	cluster 3
0	0,00%	0,00%	2,38%
1	7,45%	2,17%	9,52%
2	7,45%	6,52%	9,52%
3	13,83%	0,00%	2,38%
4	27,66%	10,87%	0,00%
5	25,53%	76,09%	40,48%
6	18,09%	4,35%	28,57%

Source: own

Tab. 17: Sex of the respondent of cluster analysis.

Sex	cluster 1	cluster 2	cluster 3
men	0,00%	2,17%	100%
woman	100,00%	97,83%	0%

Source: own

Tab. 18: Grade of cluster analysis

Grade	cluster 1	cluster 2	cluster 3
1	1,06%	2,17%	2,38%
2	94,68%	86,96%	90,48%
3	4,26%	10,87%	7,14%

Source: own

Tab. 19: Time spent at the computer of cluster analysis.

Time	cluster 1	cluster 2	cluster 3
1	22,34%	2,17%	11,9%
2	35,11%	6,52%	28,57%
3	20,21%	45,65%	21,43%
4	9,57%	26,09%	21,43%
5	0,00%	10,87%	9,52%
6	6,38%	6,52%	9,52%
7	6,38%	2,17%	9,52%

Source: own

The results of cluster analysis using K-means algorithm is little apparent differentiation between clusters. Maximum numbers of items (highlighted in the table cells) are almost identical in all properties. Therefore, for further work the results have been chosen from cluster analysis using Two-Step algorithm.

2.5 Evaluation of cluster analysis

To analyze the results of cluster analysis we chose decision trees. We tested three models of decision trees: C 5.0, Classification and Regression (C&R) Tree and QUEST.

Decision tree models allow developing classification systems that predict or classify future observations based on a set of decision rules. If we have data divided into classes that interest us, we can use our data to build rules that we can use to classify old or new cases with maximum accuracy.

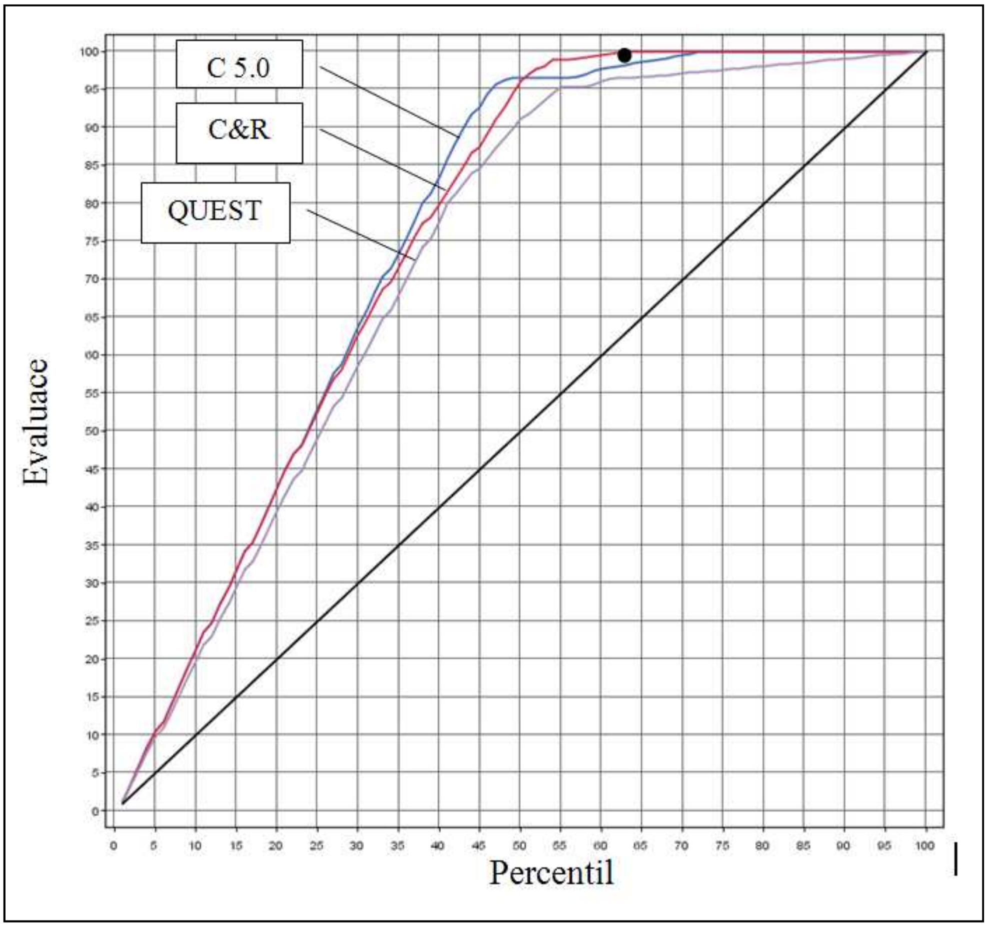
A C5.0 [8], [9] model works by splitting the sample based on the field that provides the maximum information gain. For each subsample defined by the first split is then split again, usually based on a different field, and the process repeats until the subsamples cannot be split any further.

The C&R Tree [3], [9] is a tree-based classification and prediction method. Similar to C5.0, this method uses recursive partitioning to split the training records into segments with similar output field values. C&R Tree starts by examining the input fields to find the best split, measured by the reduction in an impurity index that results from the split. The split defines two subgroups, each of which is subsequently split into two more subgroups, and so on, until one of the stopping criteria is triggered. All splits are binary (only two subgroups).

QUEST [9] is a binary classification method for building decision trees. A major motivation in its development was to reduce the processing time required for large C&R Tree analyses with either many variables or many cases. A second goal of QUEST was to reduce the tendency found in classification tree methods to favour predictors that allow more splits; that is, continuous predictor variables or those with many categories.

The target variable for the decision algorithm whether a particular item to that cluster (Two-step algorithm). The best models of decision trees were chosen based on the Evaluation (see Fig. 2) which shows that, as seems to be a model the decision tree algorithm based on algorithm C 5.0.

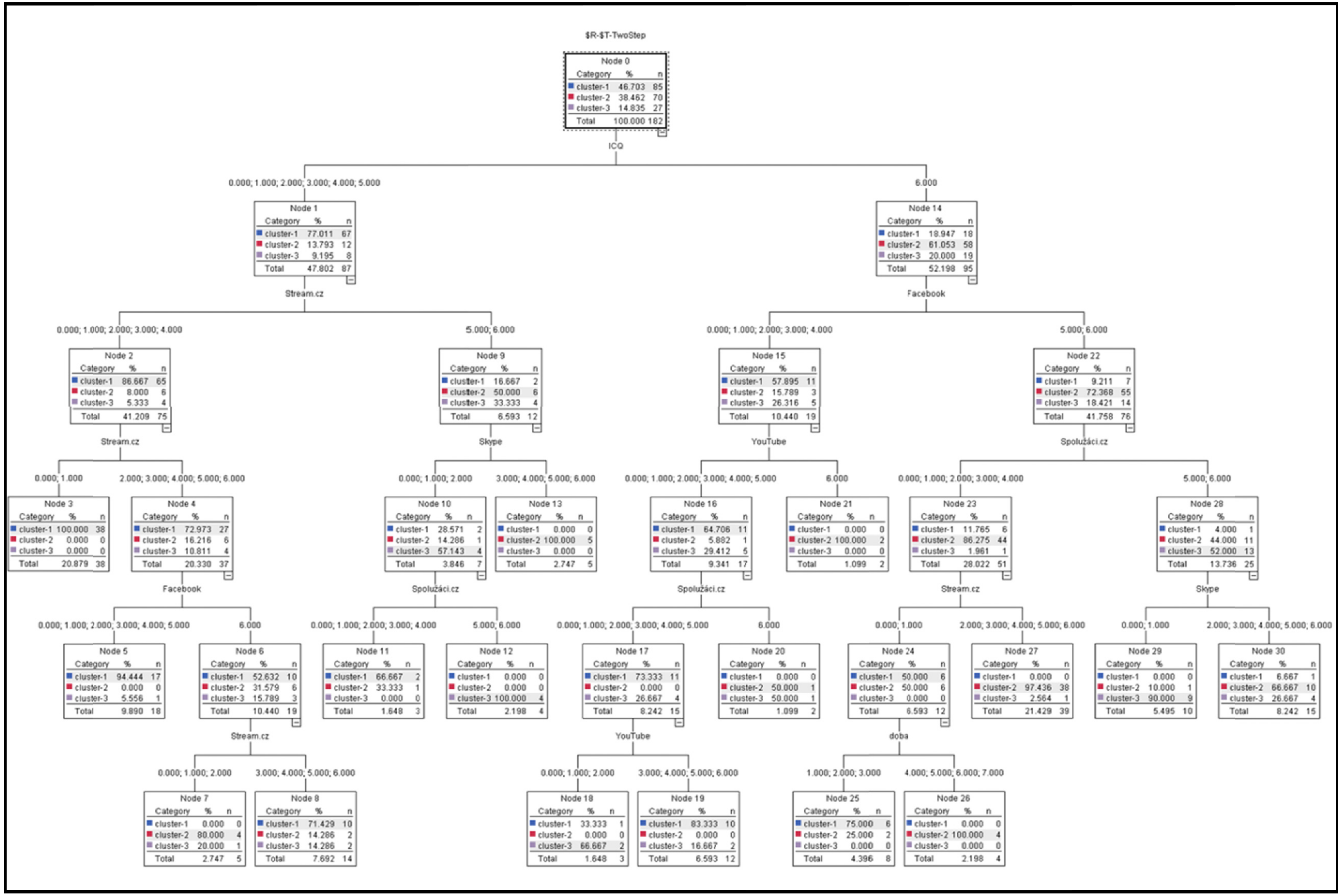
Fig. 2: Comparison of different decision algorithms with the use of the evaluation chart.



Source: own

The algorithm C&R Tree achieved 100% accuracy of classification as the one at 62% percentile. Therefore, this algorithm was chosen as a suitable explanation for the clustering analysis. The analysis resulted in the decision trees algorithm using C&T Tree is following a model (Fig. 3) that suggests how to weight the individual characteristics of the assignment to a cluster that generated the Two-Step algorithm.

Fig. 3: The decision tree generated by algorithm C&T Tree



Source: Own

Conclusion

Respondents, the students of the Faculty of Economics and Administration, can be divided into three clusters.

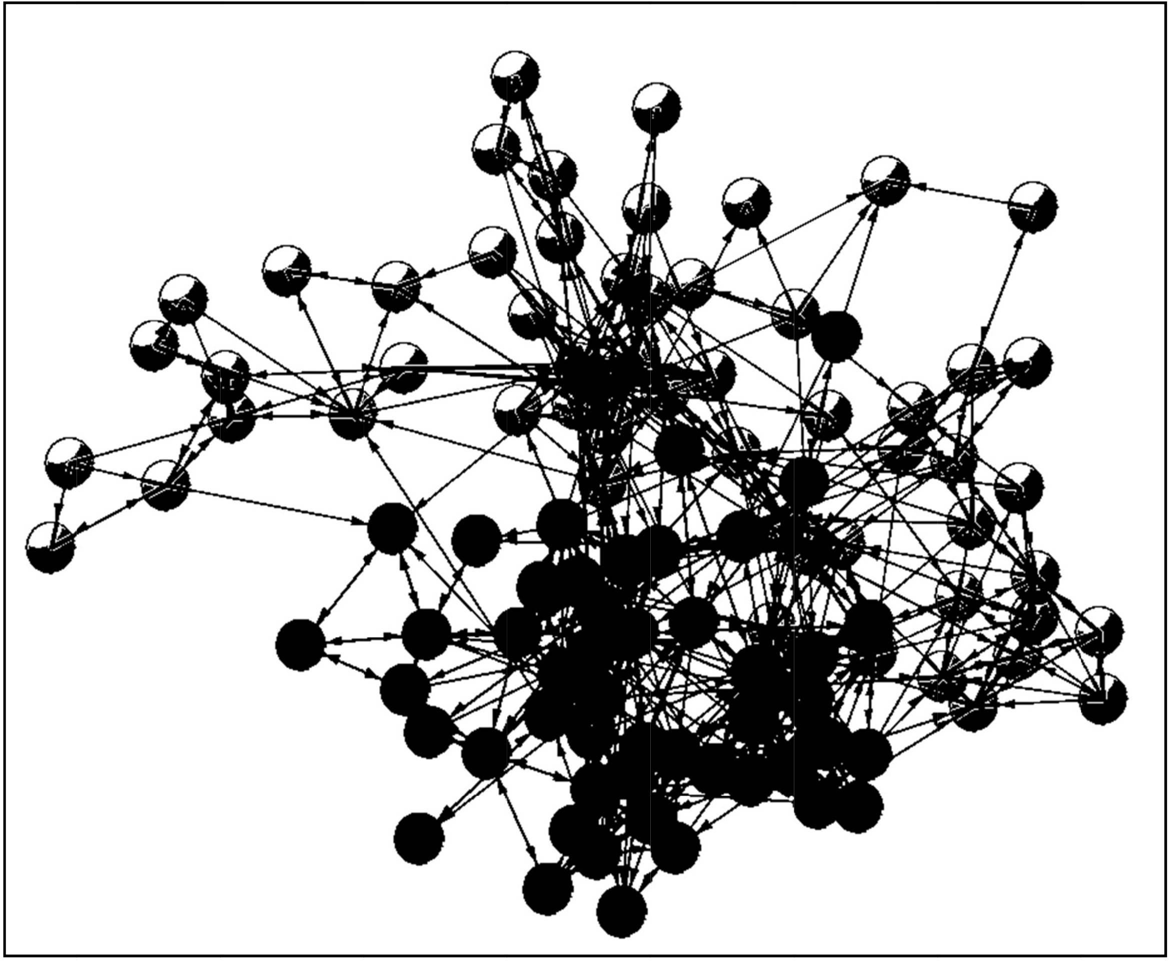
For the first cluster is characteristic that it is dependent on sex and grade. This cluster does not use any dominant communication network. Respondents prefer a communications network ICQ for daily communication. Skype is used in exceptional cases, but several times a year. However, rather consume media content on YouTube, which is used several times a week. Also this cluster is relatively latent for the Czech social network Spolužáci.cz. Students spend time at the computer between 10-20 hours per week. This cluster is also represented by largest proportion of respondents 85 (46,70%). This is a rather passive group of respondents of social networks overall.

The next cluster contains 70 respondents (38,46%), second in size. Typical for respondents in this cluster is that it is the most active group on Facebook, 85% of respondents from this cluster is on Facebook on a daily basis. Likewise, the communications network ICQ uses 82% of respondents to the daily communication. It is also the group which is very actively communicates with Skype. This cluster tends to the daily consumption of media content on YouTube. This group also spends relatively a lot of time at the computer 30-40 hours. Also in this cluster is not indicative of gender and grade. Generally, this group is described as very active on social networks.

For the last cluster, which is the smallest with 27 respondents (14,83%) is the dominant female. ICQ is very powerful tool for daily communication in this group. On other hand they almost do not use Skype. Furthermore, this group of respondents is the most active in the Czech social network Spolužáci.cz. And every day this group consumes media content on YouTube. Average time spent by this group of respondent at the computer is 35 hours per week.

Therefore, for better understanding how Generation Y communicates with each other, we plan further research to compile a real social network that will confront the type of communication a real meeting of students at the University of Pardubice and communication through computer networks. The Fig. 4 is a social network created by students themselves. Only black spots represent men and two tone (black and white) spots represent women.

Fig. 4: The social network of students who participated in the research Generation Y at the University of Pardubice.



Source: own

Number of return questionnaires were not great (approximately 30%), this was because a pilot survey. We try to motivate students to respond to both questionnaires and completing paying closer attention. The aim of this article was to describe structure of the social network of Faculty of Economics and Administration students by cluster analysis tools.

References

[1] BAATARJAV, E., PHITHAKKITNUKOON, S., DANTU, R. *Group Recommendation System for Facebook.* On the Move to Meaningful Internet Systems: 2008 Workshops: OTM 2008 Workshops, Springer-Verlag Berlin, Heidelberg, 2008. Volume 5333/2008, 211-219, ISBN 978-3-540-88874-1

[2] COVEY, N. *How Teens Use Media*. The Nielsen Company. [online] June 2009. [cit. 15.8.2010] Dostupné z www: <http://blog.nielsen.com/nielsenwire/reports/nielsen_howteensusemedia_june09.pdf>

[3] JOHANSSON, U., NIKLASSON, L., KÖNIG, R. *Accuracy vs. comprehensibility in data mining models*. Seventh International Conference on

Information Fusion: FUSION 2004. [online] [cit. 10.9.2010] Dostupné z www: <http://www.fusion2004.foi.se/papers/IF04-0295.pdf 2004>.

[4] JONÁŠOVÁ, H., MICHÁLEK, K. *Internetové sociální sítě a Generace Y na Fakultě ekonomicko-správní Univerzity Pardubice*. Scientific Papers of the University of Pardubice, Series D, Faculty of Economics and Administration. 2010, 15, 18, s. 98-107. ISSN 1211-555X.

[5] JUNCO, R., MASTRODICASA, J. *Connecting to the Net.Generation: What Higher Education Professionals Need to Know About Today's Students*. s.l. : National Association of Student Personnel Administrators, 2007. ISBN 978-0931654480.

[6] KOHONEN, T. Self-Organizing Map. New York. Springer, 2001. ISBN 3540679219

[7] MACQUEEN, J. B. *Some Methods for classification and Analysis of Multivariate Observations*. University of California Press, 1967. MR0214227.

[8] MORGAN, J. R. *C4.5: Programs for Machine Learning*. Quinlan, Kaufmann Publishers, 1993. [online] 2010. [cit. 1.10.2010] Dostupné z www: <http://www.find-docs.com/Quinlan-J-R-C4-5-Programs-for-Machine-Learning%09Morgan-Kaufmann-1993.html>

[9] SPSS . Inc., Clementine 10.1 . User Guide. SPSS Inc. 2010.

[10] STRAUSS, W., HOWE, N. *Generations: The History of America's Future, 1584 to 2069.* Harper Perennial, 1992. ISBN 978-0688119126.

Contact Address

Ing. Hana Jonášová, Ph.D.
Ing. Karel Michálek, Dis.
Ing. Jan Panuš, Ph.D.
University of Pardubice, Faculty of Economics and Administration, Institute of System Engineering and Informatics
Studentská 84, 532 10 Pardubice, Czech Republic
E-mail: hana.jonasova@upce.cz,
karel.michalek@gmail.cz
jan.panus@upce.cz
Phone number: +420 466 036 074, +420 466 036 001

Received: 30. 09. 2011
Reviewed: 02. 05. 2012
Approved for publication: 03. 05. 2012