A Data Mining Survey on Identifying the Factors that Influence Teachers' View of the Romanian Educational System

ANGEL-ALEX HĂISAN
Faculty of Sociology and Social Work
Babe□-Bolyai University
B-dul. 21 Decembrie 1989, No. 128, Cluj-Napoca
ROMANIA
haisan-angel@hotmail.com

VASILE PAUL BRESFELEAN
Faculty of Economics and Business Administration
Babe□-Bolyai University
Teodor Mihali, No. 58-60, Cluj-Napoca
ROMANIA
paul.bresfelean@econ.ubbcluj.ro

Abstract: - The evaluation of different socio-professional categories and of educational systems has been in the center of various multidisciplinary studies. In the research literature we found few surveys conducted on the teachers' quality of life. We also observed that numerous studies were performed on the evaluation of the educational systems but seen from the "exterior". In order to discover the factual situation, there ought to be conducted further studies that evaluate the educational systems from the standpoints of the ones directly involved, namely the teachers. In the current article we illustrate a part of our research on the quality of life among physical education and sport teachers from Cluj-Napoca and their view of the educational system, based on data mining techniques – decision trees for classification learning.

Key-Words: - Sociology, Data Mining, C4.5, k statistic, MAE, Physical Education and Sport Teachers, Educational System Evaluation

1 Introduction

The Romanian educational system, in the conditions of the prolonged socio-economic crisis, is confronted with a series of complex problems. According to studies made by the Institute of Science of Education supported by UNICEF, the dropout level is very high [8]. In another study entitled "School as it is" made by the Centre Education 2000+ and UNICEF, they point out the lack of professional training on the behalf of the teachers [14], which together with the absenteeism of the financial motivation have a direct influence on the poor training of the students. Another concerning aspect would be the violence in schools, as another study made by the Institute of Science of Education points out. According to this study, the number of institutions which reported acts of violence (physical, verbal, etc.) is greater than 75% [9]. All of these contribute to a decrease of the populations' perception on the educational system, also revealed by the researches completed by the Romanian Research Institute for Quality of Life regarding the quality of life in Romania [12].

The present paper represents a part of a larger study, which was made to determine the quality of life among physical education and sport teachers from Cluj-Napoca, an important city in Romania, who work in pre-university schools. In order to collect the necessary data for this study, we have distributed questionnaires to all this personnel from Cluj-Napoca. The survey has been build taking into account the indicators used by the European Quality of Life Surveys [23] in their studies of the quality of life. These indicators refer to the following aspects: economical situation, health, professional life, social life, family life, housing and environment, the level of satisfaction etc. We employ data mining techniques, such as C4.5 algorithm for classification learning, in an attempt to generate suggestive decision trees to predict the teachers' view of the Romanian educational system.

ISBN: 978-1-61804-113-5

2 Problem Formulation

The data has been collected in the scholar year 2011-2012, the questionnaires being distributed to a number of 149 teachers. Due to the fact that 44 of them declined to take part in the research, we remained in the end with a total of 105 subjects, which represent a 70,46% rate of response.

With the intention of evaluating the professional life of the teachers, we inserted in the questionnaire questions regarding: number of years of activity, regret towards their profession, evaluation of the educational system, functions and specializations held, a second place of work, justification of choosing their profession, steps in professional formation, integration in the educational system, school equipments, satisfactions offered by the job, relation between family and job, performances obtained, financial retribution and evaluation of their profession in the educational system.

After consolidating the results, in connection with the evaluation of the educational system indicator, we obtained the following: 52% of the subjects offered a <low quality> score in the evaluation of the Romanian education system, 27% declared that it was neither good nor bad, and only 14% offered a <good quality> score. 7% of the respondents have chosen not to express their opinions regarding this matter.

Taking into consideration these results, we then intended to identify the causes that led the physical education teachers to evaluate the system, of which they are a part of, as being low quality one. The importance of this indicator is high; it represents the place itself where they conduct their whole activity. Consequently, by obtaining a high percent for the <low quality> alternative of this indicator, it denoted the existence of a problem. This matter develops even more, if we consider that the value of a certain system is given, theoretically, by the sum of values of every individual belonging to that system.

Therefore, we were determined to identify the indicators having the most influence in evaluating the educational system through data mining techniques. We used data mining because it allowed us to process a large quantity of data, and along with the attributes that refer strictly to the professional life, it connected other attributes that had no visible relations with them, offering interesting patterns.

In our study, we have tried to approach the educational system evaluation from the insiders' point of view, namely the teachers, and use the data mining techniques in processing qualitative sociological data obtained from the questionnaires.

3 State of the Art Researches

The evaluation of different socio-professional categories has been in the center of various multidisciplinary studies. The Romanian Research Institute for Quality of Life studied the Romanian citizens' quality of life each year for the interval 1990-1999. These were followed by studies made prior to our country's integration in the European Union (2003 and 2006) and by one conducted in 2010. According to the last study, the Romanian citizens' quality of life has been decreasing; reports showed severe dropping for some indicators to the level of 1999 and even below those [12].

With the aim of determining the quality of life of a certain socio-professional category, a series of indicators were taken into consideration, including professional life. After consulting the research literature, we found few studies conducted on the teachers' quality of life. The most relevant one was carried out in China where the quality of life of the college teachers was studied by the Medical University from China [4]. They concluded that this indicator was below the one of the general population, principally for the reason that teachers used more time and energy to keep up with the changes in the educational system.

We also took notice that a lot of studies were carried out on the evaluation of the educational system from the "exterior", like the survey performed by the Romanian Institute for Strategies and Evaluation. According to this survey, 61% of the interviewed subjects had a low and very low opinion on the Romanian educational system [7]. This survey revealed the opinion of the general population; however we consider that it fails to expose an accurate representation of the actual situation, due to a large amount of factors influencing their opinions.

In order to discover the factual situation, there ought to be conducted more studies, that evaluate the educational system from the standpoints of the ones directly involved in the system, namely the teachers, like the study "School as it is" [14]. We consider this as being a significant advance, for the reason that problems could be identified from the inside out, based on the teachers' feedback.

The interest for physical education and sport is very low in Romania; according to the study made by the Institute of Endocrinology C.I. Parhon, approximately 60% of the Romanians have weight problems and 30% are overweight [22]. In schools, where in theory the physical culture's foundation is laid, this domain is considered a low importance one, being positioned together with music and art

classes at a certain "etc." category, as teachers affirmed.

Taking into consideration our results vis-à-vis the evaluation of the educational system from the teachers' point of view, we aimed to determine the pointers that influence their opinions, using data mining, namely decision trees. These techniques were previously used by Bresfelean to determine the students' academic failure profiles, and their will to continue their education with post-university studies [1]. Generally, data mining has been used in education to determine students' behavior and for various educational and e-learning topics [16].

These methods were connected to sociology by MIT researchers Eagle and Pentland, so as to identify the way social networks evolve in time, how predictable are people's lives and how the information flows. They named the method "reality mining" [3].

The employ of data mining in sociology has been performed with success on processing quantitative data. Mikheyenkova wrote about data mining applications on qualitative data, concluding that although these methods cannot replace the work of a sociologist, they can represent a strong base in solving various sociological problems [13].

4 Our Survey

In our research we tried to find a solution for a quality classification learning process using the attribute edu_evaluation (the evaluation of the Romanian education system) as central class. Thus, various attempts have been made using decision tree algorithms, including those presented in this paper as a result of the C4.5 algorithm. C4.5 was developed by Ross Quinlan [15] as an algorithm to generate decision trees of arbitrary depth in a top-down recursive divide-and-conquer strategy, an expansion of his prior ID3 algorithm. It also uses a "pruning" method that replaces subtrees with leaves, thus reducing overfitting [15].

We tried to point out a number of predictions on edu_evaluation attribute by generating suggestive and simple to interpret decision trees. This could be achieved under the following conditions:

1. Obtaining a superior percentage of correctly classified instances (over 50%).
2. Calculation of statistical coefficient k and obtaining a result closer to 1.
3. Calculation of Mean Absolute Error (MAE) and achieving a result close to 0.

In the following sections we make a succinct presentation of these indicators and how their calculation was prepared.

4.1 K statistic

k statistic is calculated as: [2]

$$k = \frac{P(A) - P(E)}{1 - P(E)} \tag{1}$$

Where

P(A) - the observed agreement among the coders,
P(E) - the expected agreement, the probability that the coders agree by chance.

$$k \in [-1; 1] \tag{2}$$

Where [2]

k = 1 represents perfect agreement,
k = 0 represents that agreement is equal to chance,
k= -1 represents "perfect" disagreement.

In the research literature, we could find two different methods for calculating P(E): [2]

1) each coder has a personal distribution, based on that coder's distribution of categories.
2) there is one distribution for all coders, derived from the total proportions of categories assigned by all coders.

Using the second method, in Siegel and Castellan's approach [18],[2], P(E) is calculated as:

$$P(E) = \sum_j \left(\frac{\sum_i n_{ij}}{Nk} \right)^2 \tag{3}$$

Krippendorff [10],[2] calculated P(E) (also called 1–De in his terminology) with a sampling-without-replacement methodology, but the difference is negligible with the previous approach:

$$1 - D_e = \sum_j \left(\frac{\sum_i n_{ij}}{Nk} \right) \left(\frac{[\sum_i n_{ij}] - 1}{Nk - 1} \right) \tag{4}$$

Where [2]

The agreement table = N x m matrix,
N - the number of items in the data set,
m - the number of labels that can be assigned to each object - in our example N=105 and m=4,
n_{ij} - the number of codings of label j to item i.

4.2 Mean Absolute Error (MAE)

Measures of forecast accuracy can be placed into two categories: those that are "scale-dependent" and those that are not [19]. We utilized Mean Absolute Error (MAE) as a scale-dependent measure in order to make precision assessments among data sets, so that dissimilar scales influencing the level of these indicators are not misinterpreted as discrepancies in error.

Individual model-prediction errors are usually defined as: [20]

$$e_i = P_i - O_i, \quad i = 1,2,...,n \tag{5}$$

Where

Pi - predictions or statistical comparisons of model estimates

Oi - pair wise matched observations.

Measures of average error or model performance are based on statistical summaries of ei (i = 1, 2,...,n). Accordingly, the Mean Absolute Error (MAE) represents the average of the absolute values of the differences between predictions and the corresponding observation, and was calculated as follows: [20]

$$MAE = n^{-1} \sum_{i=1}^{n} | e_i | \tag{6}$$

MAE is a linear score which signifies that all the individual differences are weighted uniformly in the average [5]. It is also a negatively-oriented score, ranging from 0 to ∞, that is why we sought out lower MAE values (lower values are considered to be better).

4.3 Decision tree

In Table 1 we present a part of the experiment based on C4.5 algorithm in RapidMiner and Weka software [21]. There have been numerous attempts, using different values of the parameters employed in C4.5, such as:

- Min_NumObj - minimum number of instances per leaf.
- Num_Folds - the amount of data used for reduced-error pruning. One fold is used for pruning, the rest for growing the tree.
- Pruned - whether the process of pruning is performed.
- Laplace - whether counts at leaves are smoothed based on Laplace estimator (which initiates all numbering starting with 1 as a substitute of 0).

Table 1. Parameters and indicators in the C4.5 decision trees experiments

num_Folds	min_NumObj	Leaves No.	Tree Size	Pruned/ Laplace	Correctly Classified Instances%	Correctly Classified Instances	Kappa statistic	Mean absolute error
1	1	52	63	Pruned / Laplace	75.2381 %	79	0.5483	0.1838
1	1	140	171	UnPruned / NoLaplace	100%	105	1	0
2	2	77	94	Pruned / Laplace	85.7143 %	90	0.7581	0.2728
2	4	11	14	Pruned / Laplace	63.8095 %	67	0.3375	0.2875
2	7	17	21	UnPruned / NoLaplace	62.8571 %	66	0.3166	0.2361
3	2	20	25	Pruned / NoLaplace	57.1429 %	60	0.1834	0.3016
3	9	18	20	UnPruned / Laplace	60%	63	0.2365	0.2996
3	*14*	*11*	*14*	*UnPruned / Laplace*	*69.5238 %*	*73*	*0.4346*	*0.2237*
4	5	11	14	Pruned / Laplace	63.8095 %	67	0.3375	0.2875
4	6	17	21	UnPruned / Laplace	62.8571 %	66	0.3166	0.2856
4	8	1	1	Pruned / Laplace	52.381 %	55	0	0.317
4	10	18	20	Pruned / Laplace	60%	63	0.2365	0.2996
5	1	140	171	UnPruned / Laplace	100%	105	1	0
5	4	31	40	UnPruned / NoLaplace	73.3333 %	77	0.5553	0.1715
5	8	15	18	UnPruned / Laplace	62.8571 %	66	0.3166	0.288
5	11	18	20	UnPruned / NoLaplace	60%	63	0.2365	0.2566
5	12	1	1	UnPruned / Laplace	52.381 %	55	0	0.317

We obtained decision trees of different sizes (Leaves No. and TreeSize parameters), with a lesser or greater percent of Correctly Classified Instances. In order to validate our experiment, there have been calculated several statistical indicators, from which we present the results of K statistic and MAE.

In choosing the most suggestive tree, we relied on these indicators, and also on the Correctly Classified Instances, so as the tree to be suggestive and easy to interpret. Thus, we preferred a decision tree with TreeSize=14 si Leaves No.=11, and a percentage of Correctly Classified Instances =

69.5238%, compared to a tree with 100% Correctly Classified Instances, but TreeSize = 171 and Leaves No.== 140.

Importantly, as we mentioned, was and the computation of K statistic for which we pursued a value close to 1, and MAE where we pushed for a value close to 0. Thus, for the selected tree, we obtained:

- Kappa statistic = 0.4346
- Mean absolute error = 0.2237

In Fig. 1, we present the generated decision tree in graphical form:

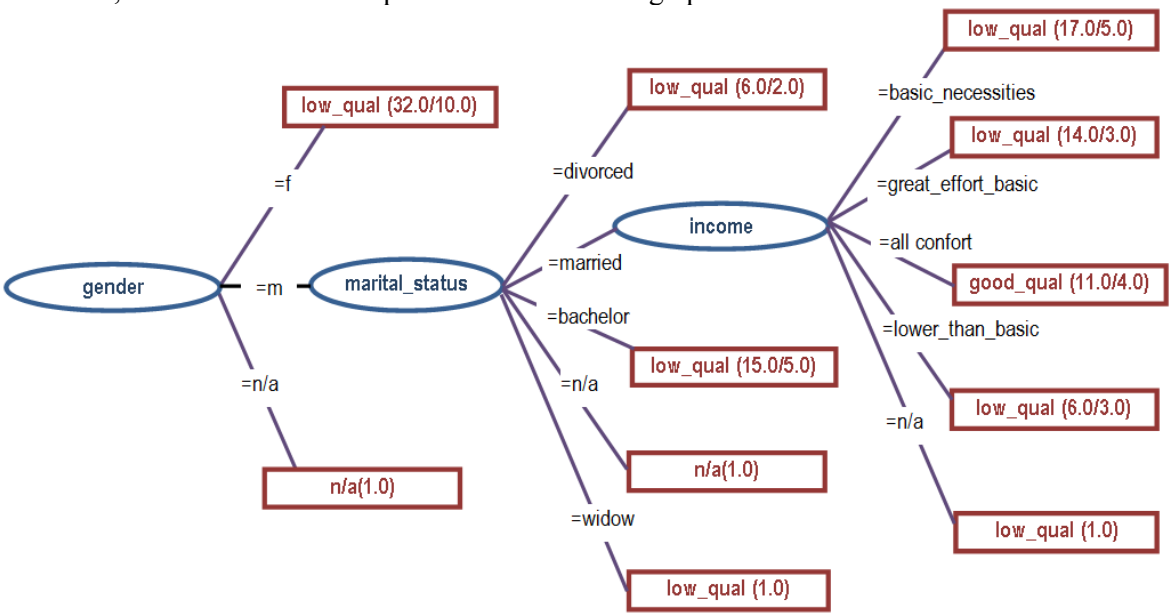

Fig.1 Selected C4.5 decision tree

Here are some evocative samples of interpretation of the decision tree's branches:

"If the teachers' gender was female, then they would offer a <low quality> score in the evaluation of the Romanian education system."

"If the teachers' gender was male, and their marital status was <married>, and their current profession income covered only the basic necessities, then they would offer a <low quality> score in the evaluation of the Romanian education system."

"If the teachers' gender was male, and their marital status was <married>, and their current profession income provided them with all the comfort, then they would offer a <good quality> score in the evaluation of the Romanian education system."

"If the teachers' gender was male, and their marital status was <bachelor>, then they would offer a <low quality> score in the evaluation of the Romanian education system."

5 Conclusion

Following a part of our experiments and based on the selected decision tree, we can conclude that the evaluation of the educational system done by the physical education teachers is influenced by three factors: gender, marital status and income. The female teachers mostly consider that the educational system has a low quality, without any influence by other factors than gender.

In the case of male teachers, the situation is slightly different. We observe that family plays an important role in the teachers' opinions about the educational system. The ones that are divorced, single and widow offer a <low quality> score in the evaluation. The married ones could be influenced by the income indicator, mainly because the necessities of a family are greater than the ones of an individual. It can also be tied to the findings from another part of our research, in which we discovered that most of them still belonged to the traditional family model, where the father had a great financial responsibility – being considered the main provider. We can observe that subjects with lower income than the basic necessities, the ones with income barely covering basic necessities and the ones that manage with great efforts, mainly consider that the educational system has a low quality. Only the teachers who manage to have everything they need and are also married, could offer a <good quality> score in the systems' evaluation.

In conclusion, we can affirm that in this phase of our research, data mining aided to identify some of the most important factors in teachers' own evaluation of the educational system. Although it

ISBN: 978-1-61804-113-5

scarcely explains the whole relationships between the indicators, necessitating a thorough analysis by sociologists, it offers a starting point in the study of unapparent pattern identification and for our future research.

In the continuation of our studies we desire to determine the part played by other factors in the teachers' evaluation of the Romanian Educational System, as well as to develop the research through other sociological and data mining methods.

Acknowledgement

The practical research of this paper was partly supported by CNCSIS TE 316 Grant.

References:

[1] Bresfelean V.P., *Implicații ale tehnologiilor informatice asupra managementului instituțiilor universitare,*[Implications of Information Technology on Universities' Management] Ed. Risoprint, Cluj, 2008.

[2] Di Eugenio B., Glass M., The Kappa statistic: a second look, *Computational Linguistics*, Vol.30, Issue 1, MIT Press, 2004, pp. 95-101

[3] Eagle N., Pentland A., *Reality mining: sensing complex social systems*, Springer-Verlag, 2005.

[4] Ge C. et al., Quality of life among Chinese college teachers: A cross-sectional survey, *Public Health*, Vol. 125/ 5, 2011, pp. 308-310.

[5] Hematinezhad M., Ramezaniyan M.R., Gholizadeh, M.H., Shafiee, S.H., Ghazi Zahedi, Predicting National Team Rank in Asian Game Using Model Tree, *Pamukkale Journal of Sport Sciences*, Vol. 2, No. 3, 2011, pp.22-36.

[6] Hsu C-H, Lee T-Y, and Kuo H-M. Mining the body features to develop sizing systems to improve business logistics and marketing using fuzzy clustering data mining. *WSEAS Transactions on Computers*. 8, 7 (July 2009), pp.1215-1224.

[7] IRES, *Percep□ii privind sistemul educa□ional din România* [Perceptions on Romanian educational system], Raport de cercetare - Sondaj de opinie, 2010 –2011, Bucuresti, 2011.

[8] Jigau M. et al., *Estimarea dimensiunii fenomenului de abandon scolar folosind metodologia analizei pe cohorta*, [Estimating the size of drop out using the cohort analysis methodology]. Bucuresti, 2011. www.oportunitatiegale.ro/pdf_files/ise-unicef%20studiu-abandon.pdf

[9] Jigau M. et al., *Violenta in scoala* [Violence in school], Alpha MDN, Bucuresti, 2006.

[10] Krippendorff, K., *Content Analysis: an Introduction to its Methodology*, 1980, Sage Publications.

[11] Mamcenko J. and Beleviciute I., Data mining for knowledge management in technology enhanced learning. In *Proceedings of AEE'07*. WSEAS, pp. 115-119

[12] Marginean I. et al., *Calitatea vietii in Romania 2010*, ICCV, Bucuresti, 2010.

[13] Mikheenkova M. A., Computer-Support Capabilities for Qualitative Research in Sociology, *Automatic Documentation and Mathematical Linguistics*, Vol. 45, No. 4, 2011, pp. 180–201.

[14] Nedelcu A. et al., *□coala a□a cum este* [School, the way it is], Centrul Educa□ia 2000+, Unicef, Bucuresti, 2010.

[15] Quinlan JR. *C4.5: Programs for Machine Learning*. Morgan Kaufmann, 1993

[16] Romero C., Ventura S., Educational Data Mining: A Review of the State-of-the-Art, *Transactions on Systems, Man, and Cybernetics--Part C: Applications and Reviews*, Vol. 40 Issue 6, November 2010

[17] Saitta S., Raphael B., Smith I.F.C., Data Mining for Decision Support in Multiple-Model System Identification, *Proceedings of SMO'06*, WSEAS, 2006, 161-166

[18] Siegel S., Castellan N.J. Jr., *Nonparametric statistics for the behavioral sciences*, 1988, McGraw Hill.

[19] Swanson D.A., Tayman J., Bryan T.M., MAPE-R: a rescaled measure of accuracy for cross-sectional subnational population forecasts, *Journal of Population Research* 04 / 2012, 28(2), pp.225-243.

[20] Willmott C.J., Matsuura K., Advantages of the mean absolute error (MAE) over the root mean square error (RMSE) in assessing average model performance, *Climate Research*, Vol. 30, 2005, pp. 79–82.

[21] Witten I.H., E. Frank, Hall M.A*., Data mining : practical machine learning tools and techniques*.-3rd ed., Elsevier, 2011

[22] ***Institutul de Endocrinologie C.I. Parhon, *Mai mult de jumatate dintre romani au probleme cu greutatea* [More than half of the Romanians have weight problems],AgerPress, 2009, www.ziare.com/social/capitala/mai-mult-de-jumatate-dintre-romani-au-probleme-cu-greutatea-818977

[23] ****European Quality of Life Surveys* – EQLS, www.eurofound.europa.eu/surveys/eqls/index.htm