Classification methods as a tool for discovering the attributes with influence on modelled socio-economic characteristics in the studied region

Michael ROST

Pavel TLUSTÝ

Renata KLUFOVÁ

Dagmar ŠKODOVÁ-PARMOVÁ

Abstract

The aim of the paper is to show some modern or not so typical classification methods as sophisticated tools for identification of the factors influencing some socio-economic factors in the analyzed micro-region. For this purpose, we used statistical methodology like classification and regression trees (CART) and Chi-square automatic interaction decision method (CHAID). These classification methods are not parametric methods and their application is not restricted by a strong normality assumption. Our approach is demonstrated on one region in the Czech Republic. Data pre-processing and the numerical computation were carried out in the programming language R.

Key words:

Modelling, non-parametric classification methods, region, socio-economic attributes

Introduction

It is very important to understand the mechanism of the discrepancy of development in different rural areas. Such discrepancy is usually expressed by different dynamics of various socio-economic attributes like income of the area, number of entrepreneurs in the particular area, migration, etc. One of the big contemporary problems is transmigration of the young generation between aged between 18 and 35 into towns.

In this contribution, we focus on the willingness of people to stay to live in specific micro-region in South Bohemia. We consider the willingness to stay to live in the area to be one of the main pre-determinants to sustainable development of an area. Our aim is to enlarge the results obtained by Rost, Cermakova, Biskup [4].

Material and Methods

The analyzed subset of the data contains results of a small opinion poll survey in the chosen municipalities of the mountain and sub-mountain micro-regions in the Region of South Bohemia. More concretely, we chose the micro-region of Kaplicko, which could be characterized as a mountain and sub-mountain micro-region in the district of Český Krumlov

close to the Czech-Austrian borders. Our work was focused on young people aged 18 to 35. We divided this group into three exclusive categories: [18-24], [25-29] and [30-35] and named them as "early juvenility", "juvenility" and "almost middle age". People were asked about the place for life in the future. The respondents could answer the posed question using the following statements: {I stay here, I move into another municipality, I will move to town, I don't know}.

The following attributes were hobbies of respondents: walking, do-it-yourself work, gardening, hunting, fishing, farming, sport, culture, etc. At the same time, we recorded the number of children {0, 1, 2, 3 and more} and education of particular respondents {elementary school, vocational school, vocational school with secondary school leaving examination, secondary, university}.

To determine main factors influencing migration of inhabitants from particular area, we used nonparametric classification methods like classification and regression trees based on approach of Breiman [1] and on CHAID method proposed by Kass [3]. We can briefly describe these methods as follows:

CART like the others methods (C4.5, C5.0 for example) is a tree-based method which partition the feature space into a set of rectangles. During the building the set of classification rules, e.g. tree growing, we usually employ the following phases [4]:

I. The split criterion for each node of growing tree is chosen. This problem is usually solved by impurity measure. As an impurity measure, we chose the Gini index. Other possibilities are for example Misclassification error, Cross-entropy or deviance for more technical details see [1] or [2]. The Gini index could be expressed as:

$$i(t) = \sum_{i \neq j} p(i \mid t)\, p(j \mid t)\,,$$

where $p(j|t)$ stands for probability that object is in the node t and is assigned to class j. Computation of this probabilities is based on a priory selected set of objects, this set is usually called as training data set.

II. To decide which node becomes a leaf (terminal node of tree) is an essential problem solved in the second stage. Usually pruning of tree is used. After building T_{max} tree (each leave contains the objects only from one class, or the number of classifying objects in each leave is smaller than the prescribed value) is this tree pruned to tree T_{optim}. The new tree T_{optim} is the subset of T_{max}. This pruning of tree T_{max} to tree T_{optim} minimize estimation of relative error of classification $R(T)$.

III. The third phase is the simplest part of the tree growing process. Each of the classes is assigned to one of the leaves. The idea is following: to assign correctly the specific class to leaves t is to assign the value, that minimizing the estimate of the misclassification error, e.g. to assign the value satisfying the following equation:

$$p(j(t) \mid t) = \max_j \frac{N_j(t)}{N(t)}$$

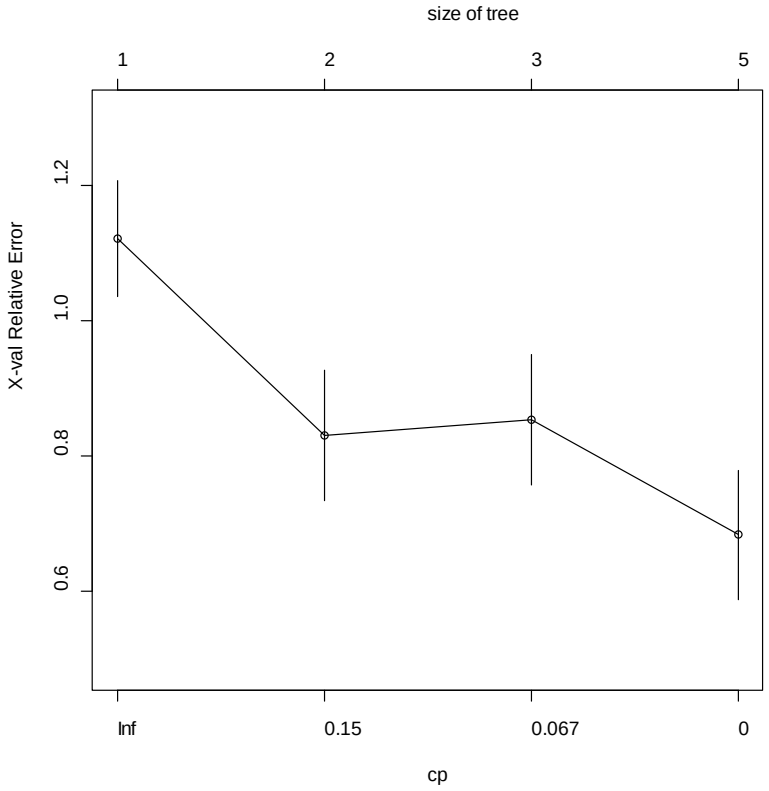
More information about CART methodology can be found in [1] or in [2]. As we can see, the major advantage of the recursive binary tree is its nice interpretability. The whole feature space partition is fully described by one tree.

The second used method was CHAID method. CHAID stands for acronym: χ^2 Automatic Interaction Detection proposed by Kass [3]. It detects interactions between categorized variables of a data set, one of which is the dependent variable. CHAID algorithm is based on maximizing the significance of a chi-squared statistic for cross-tabulations between the dependent variable and the predictors at each partition. The data are partitioned into mutually exclusive, exhaustive subsets that best describe the dependent variable. On the opposite to CART the CHAID permits non-binary splits [3].

Whole data preprocessing and the numerical computation and model building were carried out in the programming language R with use of libraries called rpart [5] and CHAID [6].

Results

At the beginning of the building classification rules, the sufficiently branched tree T_{max} was created. To secure the growing process, the complexity parameter *cp* was specified by zero value because the low value of the complexity parameter made the tree sufficiently branched. To decide where stop the growing process and to prune the tree, we used 1-SE rule. For technical details of pruning process see [1]. Graph 1 shows that tree with size 5, e.g. with five terminal nodes - leaves satisfies this rule. So we could not prune the grown tree anymore.

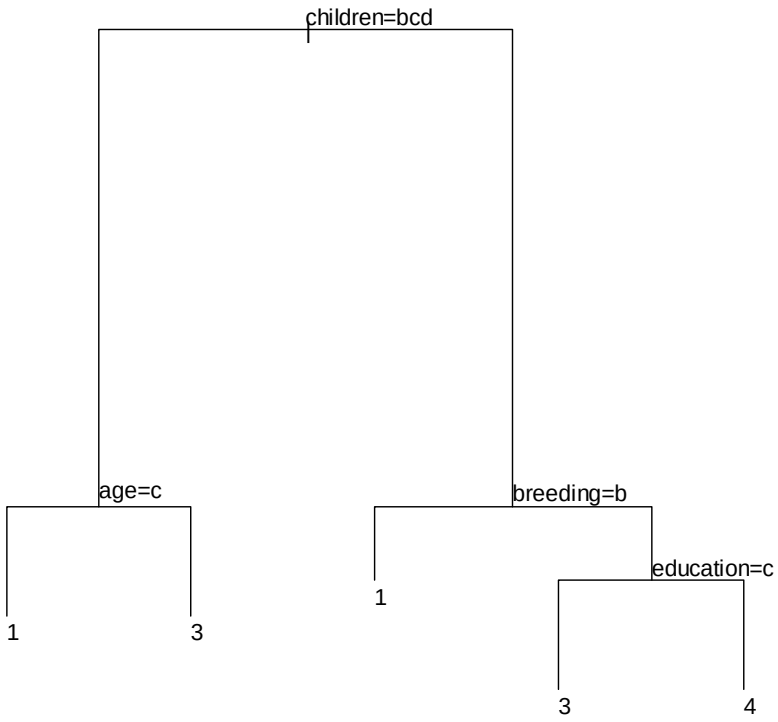


Corresponding classification rules obtained from this process are shown in Table 1. We can see that in the tree growing process, only the four attributes were identified as significant factors. These are: number of children, age of respondent, education of respondent and farming.

Type of node	Node	Classification rule	Number of objects	Classification[a] [no. of misclassification]	Conditional probabilities			
					(t|1)	(t|2)	(t|3)	(t|4)
Root	1)		63	1 [41]	0,3492	0,0476	0,3175	0,2857
Node	2)	Children 1 or 2 or 3	27	1 [10]	0,6296	0,0741	0,2963	0,0000
Leaf	4)	Age = S	15	1 [2]	0,8667	0,0667	0,0667	0,0000
Leaf	5)	Age= HM or SM	12	3 [5]	0,3333	0,0833	0,5833	0,0000
Node	3)	Children = 0	36	4 [18]	0,1389	0,0278	0,3333	0,5000
Leaf	6)	Farming = 1	7	1 [3]	0,5714	0,0000	0,1429	0,2857
Node	7)	Farming = 0	29	4 [13]	0,0345	0,0345	0,3793	0,5517
Leaf	14)	Education =SS a VS	11	3 [5]	0,0909	0,0909	0,5455	0,2727
Leaf	15)	Education =SOU a SOUM	18	4 [5]	0,0000	0,0000	0,2778	0,7222

Classification[a]: 1 – stay in current area, 2 – Move to another municipality (other countryside), 3 – move to the city (town), 4 – not decided move or stay in the area.

Graphical representation of the classification rules are in Graph 2 in form of classification trees. Observation satisfying the condition at each junction is assigned to the left branch and the others to the right branch.



The main conclusion could be formulated as follows. Childless respondents who are not interested in farming, with secondary or university education, will move to towns. Respondents with elementary school and those who are not interested in farming do not think about the posed question. On the other hand, respondents with one or more children (respondents with more than 3 children were not included) and respondents aged 30-35 will not move out of the studied micro-region. Hobbies and education have no influence on this decision in this specific group of respondents. The same is true for childless respondents who like farming.

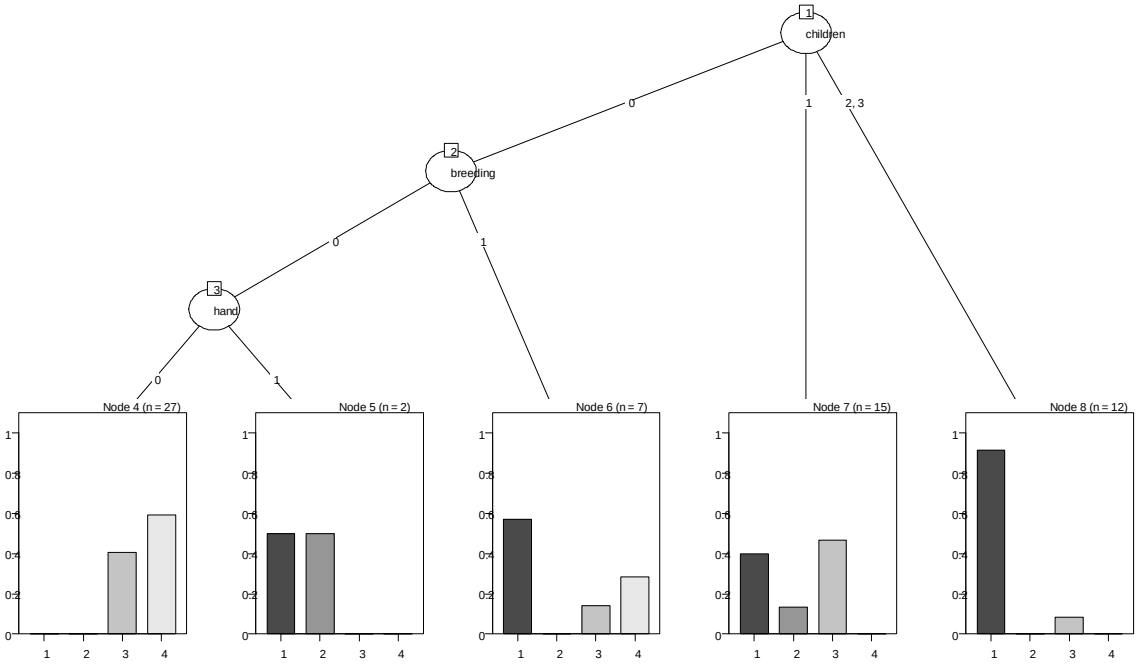
In case of CHAID method, we discovered only three significant factors, but during the construction process, we set the parameter alpha (for splitting) to 0.025. We chose this value with aim to obtain classification tree comparable to tree grown with CART methodology, e.g. the classification tree with the five terminal nodes - leaves. In this case significant factors are: number of children, farming and possibility of manual working (DIY). The derived classification rules are provided in Table 2. The number of misclassification and the error rate is derived only for the terminal nodes, e.g. leaves of the constructed classification tree.

Type of node	Node	Classification rule	Number of objects	Classification[a] [no. of misclassification, error in %]
Root	1)	----	63	- [--,error = --]

Node	2)		Children = 0	--	- [--,error = --]
Node		3)	Animal husbandry = 0 (no)	--	- [--,error = --]
Leaf		4)	Hand = 0 (no)	27	4 [11, error = 40.7 %]
Leaf		5)	Hand = 1 (yes)	2	1 [1, error=50 %]
Leaf		6)	Animal husbandry = 1 (yes)	7	1 [3, error=42.9 %]
Leaf	7)		Children = 1	15	3 [8, error=53.3 %]
Leaf	8)		Children = 2 or 3		1[12, error= 8.3 %]

Classification[a]: 1 – stay in current area, 2 – Move to another municipality (other countryside), 3 – move to the city (town), 4 – not decided move or stay in the area.

For better understanding of the classification rules, we provided also the classification tree using CHAID methodology (see Graph 3). It is apparent that in this case, the first split was not binary. For a clearer picture of the classification results, see the distribution of classes in particular leaves which is provided using bar plots in the terminal nodes of the tree.



The quality of the classification rules obtained using both methods was evaluated by the resubstitution method. The results are presented in Table 3. In case of CART methodology, we obtained classification efficiency of 68.25 %. The classification efficiency reached by CHAID algorithm was a slightly lower with comparison to CART. More concretely, this approach reached 61.91 %. For more specific results, see Table 3.

	Continuance in the same municipality – stay	Move to municipality (other countryside)	Move to the city (town)	Do not care about it / do not know
Continuance in the same municipality - stay	17[a]; 16[b]	0[a]; 0[b]	5[a]; 6[b]	0[a]; 0[b]
Move to another municipality (other countryside)	1[a]; 1[b]	0[a]; 0[b]	2[a]; 2[b]	0[a]; 0[b]
Move to the city (town)	2[a] ; 2[b]	0[a]; 0[b]	13[a]; 7[b]	5[a]; 11[b]
Do not care about it / do not know	2[a]; 2[b]	0[a]; 0[b]	3[a]; 0[b]	13[a]; 16[b]

[a]Results for CART; [b]Results for CHAID

Discussion and conclusion

We used nonparametric classification methods such as CART and CHAID methodology to analyze primary data obtained from respondents living in the sub-mountain micro-region of Kaplicko with regards to possibility of the future migration of inhabitants from this area. Using the above mentioned methodology, we identified some essential factors influencing the willingness to stay in the studied area. These factors include number of children, age of respondents, education of respondents, possibility of farming and possibility of manual work (DIY). Therefore municipality representatives and local management should provide good conditions for families with children and create some incentives for people with higher education. We can say that the knowledge of these pre-determinant factors could help to regional or municipal management to build a right system of invitation during formulations of strategies how to increase the number of the permanent inhabitants in the specific area.

Special thanks

This work was supported by RP 6/8 "International collaboration of the USB".

Literature

[1] Breiman, L., Friedman, J.H., Olshen, R. A., Stone, Ch. J. (1998) Classification and Regression trees, Chapman & Hall/CRC, Boca Raton, 359.p., ISBN: 0-412-04841-8

[2] Hastie T., Tibshirani, R., Friedman, J.(2001) The Elements of Statistical Learning, Data Mining, Inference and Precision. Springer, New York, ISBN 0-387-95284-5

[3] Kass G. H. (1980). An Exploratory Technique for Investigating Large Quantities of Categorical Data. Applied Statistics, 29(2), 119–127.

[4] Rost, M., Čermáková, A., Biskup, R. (2007) Determination of main factors influencing willingness to live in micro-region of Kaplicko, Lucrari scientifice Management Agricol, Editura agroprint Timisoara, Seria I, Vol. IX (2), p. 149 - 156, ISSN 1453-1410

[5] Therneau T. M., Atkinson B, R port by Brian Ripley (2011). rpart: Recursive Partitioning. R package, version 3.1-50. Retrieved October 20, 2011, from http://CRAN.R-project.org/package=rpart

[6] The FoRt Student Project Team (2009). CHAID: CHi-squared Automated Interaction Detection R, package version 0.1-0., Retrieved October 20, 2011, from https://r-forge.r-project.org/bin/windows/contrib/latest/CAHID_0.1-0.zip

Author's address:

Ing. Michael Rost, Ph.D., Department of Applied Mathematics and Informatics, Faculty of Economics, University of South Bohemia in České Budějovice, rost@ef.jcu.cz

Prof. RNDr. Pavel Tlustý, CSc., Department of Applied Mathematics and Informatics, Faculty of Economics, University of South Bohemia in České Budějovice, tlusty@pf.jcu.cz

RNDr. Renata Klufová, Ph.D., Department of Applied Mathematics and Informatics, Faculty of Economics, University of South Bohemia in České Budějovice, klufova@ef.jcu.cz

Ing. Dagmar Škodová Parmová, Ph.D., Department of Structural Policy of the EU and Rural Development, Faculty of Economics, University of South Bohemia in České Budějovice, parmova@ef.jcu.cz