Proceedings of the 7th Symposium FENS, Lublin, May 14–17, 2014

Q-Entropy Approach to Selecting High Income Households

K. Gajowniczek, K. Karpio, P. Łukasiewicz, A. Orłowski, and T. Ząbkowski

Faculty of Applied Informatics and Mathematics, WULS-SGGW, Nowoursynowska 159, 02-776, Warsaw, Poland

A generalized algorithm for building classification trees, based on Tsallis q-entropy, is proposed and applied to classification of Polish households with respect to their incomes. Data for 2008 are used. Quality measures for obtained trees are compared for different values of q parameter. A method of choosing the optimum tree is elaborated.

DOI: 10.12693/APhysPolA.127.A-38

PACS: 05.45.Tp, 05.40.Ca, 07.05.Kf, 07.05.Mh

1. Introduction

Investigations of incomes are very important for a variety of reasons. Low incomes studies, being a part of poverty analysis, are essential from a state social policy. It should be mentioned that studies of poverty go beyond the analysis of income distributions. They also include an analysis of the characteristics of households belonging to the realm of poverty or being in danger of the social exclusion [1, 2]. On the other hand, the highest incomes often arouse emotions as they can imply high social inequalities. From the point of view of governments the high incomes are important because of the fiscal and social transfers (income redistribution). The highest incomes distributions are rather specific. It is clearly visible in the case of individual incomes, where we see an exponential distribution for approximately 95% of the incomes, and power law distribution for the top 5% of incomes. This was shown to be true for USA and UK [3], Australia [4], and UE [5, 6]. It also turns out that models known in the economics of income distributions, i.e., Dagum distribution and Singh-Maddala distribution [7], do not explain the highest income, despite being characterized by high compatibility and thick tails [8]. As in the case of poverty, it is also interesting to study the highest revenue from the viewpoint of the characteristics of households. It is known that certain characteristics of households, i.e., education, place of residence, or sex, have an impact on the amount of income. It is interesting to find the hierarchy of attributes that are related to the high-incomes and to show their interrelationships. One of the tools to recognize these features are classification (or decision) trees. Such methods are increasingly valued in the area of knowledge discovery (data mining), mainly due to a simple hierarchical structure and the ability to design clear rules for classification.

In economics, classification trees are widely used in marketing, logistics, banking, and credit risk [9–11]. In the field of income analysis, these methods are not widespread. The authors have used the SQL Server Analysis Services and the Statistical Analysis System (SAS) for the classification of households due to the amount of income [12]. This work is a continuation of our previous studies. We propose a generalized tree classification algorithm based on the Tsallis q-entropy. We consider two classes of income: HIGH and REST. We discuss measures of the quality of the trees for some range of the parameter q. We present a method to choose the optimal tree, which will have both high quality and relatively low, acceptable structure. For the selected tree we show the characteristics of households with the highest incomes.

2. Classification trees

Classification trees are one of the methods of multidimensional data analysis, the beginnings of which were around 60'ties of the XX century [13]. A very fast development of algorithms used in classification trees took place in eighties and nineties [14, 15]. Nowadays, classification trees are widely used and still developing.

Classification trees are an example of statistical learning methods. One randomly chooses learning sample from a set of objects characterized by independent variables (attributes). Values of dependent variable (classes) must be known for each selected object. A hierarchy of attributes is determined and rules of splitting objects among subsets of homogeneous class composition are being set out. Based on results of the calculations a tree is constructed and its parameters are evaluated. The hierarchical structure is created, which is often presented graphically as an inverted tree with a root, nodes and leaves (terminal nodes). Each node and leaf has the assigned class based on the decision threshold, usually set to 0.5. In other words, standard decision threshold choses class with bigger share. Then, the tree is being tested on a separate data set. The tree is pruned in order to exclude the most specific rules and its accuracy is being evaluated.

In this paper we assume that objects belong to two given classes (HIGH and REST) and use a classic algorithm C 4.5 [15] to construct a binary tree. Let m be a measure of diversity of objects in a given node. Let assume that the node N is divided into nodes N_1 and N_2. Then

$$\frac{|N_1|}{|N|}m(N_1)+\frac{|N_2|}{|N|}m(N_2) \qquad (1)$$

is a joint diversity of objects in both N_1 and N_2. Consequently

$$Gain(N\to N_1,N_2)=$$
$$m(N)-\left(\frac{|N_1|}{|N|}m(N_1)+\frac{|N_2|}{|N|}m(N_2)\right) \qquad (2)$$

is a corresponding information gain. The gain is being maximized in the following manner. For a given attribute one considers all divisions of the attribute into two disjoint subsets. The procedure is repeated for every attribute. A division yielding to the biggest information gain is selected. If there is no divisions related to positive information gain the node becomes leave.

Often used diversity measure is the Shannon entropy [16]

$$H = -\sum\nolimits_{i=1}^{k} p_i \log_2 p_i, \tag{3}$$

where $0 < p_i \le 1$ and $p_1 + p_2 + ... + p_k = 1$.

There are two interesting parametric generalized entropies: Tsallis [17] and Renyi [18]. Both were used in the C4.5 decision trees' algorithms, compared to each other and discussed in the paper [19]. In this work, for technical reasons related to the implementation factors, we focus on the Tsallis entropy defined as

$$H_q = \frac{1 - \sum_{i=1}^{k} p_i^q}{q - 1}, \tag{4}$$

where $p_1 + p_2 + ... + p_k = 1$ and q is a real number. Our preliminary investigations suggest that results obtained via Renyi entropy would be qualitatively similar to those following from the Tsallis entropy. A detailed comparison of the Tsallis and Renyi entropic measures in the context of income classification will be performed elsewhere. Tsallis entropy recovers Shannon entropy when $q \to 1$. If objects belong to two classes with probabilities of p and $1 - p$, we have

$$H = -(p \log_2 p + (1 - p) \log_2 (1 - p))$$

$$\text{and } H_q = \frac{1 - p^q - (1 - p)^q}{q - 1}. \tag{5}$$

The diversity measure of objects in the nodes we define as follows:

$$m = \begin{cases} H & \text{for } q = 1 \\ H_q & \text{for } q > 0 \ \& \ q \neq 1. \end{cases} \tag{6}$$

We use this measure in the further analysis. We limit q to positive values so the measures are zero if all objects in a node belong to the same class and are maximal if objects are equally distributed among classes. The measures for various values of q are plotted in Fig. 1.

3. Data

In these studies micro-data regarding budgets of households have been analyzed for year 2008, the newest complete data set available to us at the time of preparing the paper. We studied also the data for years 2000–2007 and the results were consistent. The data were collected within the project Household Budget Survey [20] and consisted of 37 107 households. The households were classified based on their 10 attributes (independent variables) belonging to the three groups: 1. variables describing a head of the household (a person with the biggest income); 2. variables describing a household as a whole; 3. variables describing a location of the household. All the attributes and their possible values are summarized in the Table I. We only point out at this place that the socio-economic group is defined as the main source of income of the household. A majority of households are employee's households (about 50%). On the other hand the smallest group consists of households maintained from non-earned sources (about 3.5%).

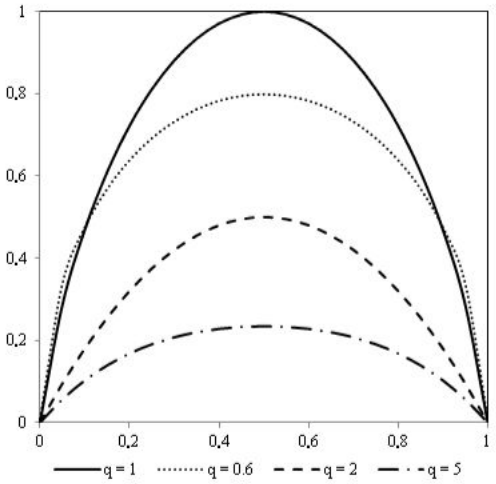

Fig. 1. Shannon and Tsallis entropies for two-point distribution.

We study household's annual income per number of earners. Taking into account a standard decimal distribution we consider two groups of households: HIGH (10% of households with the highest incomes) and REST (remaining households). The chosen value of 10% assures us the sufficiently large number of HIGH objects. The results are quite robust with respect to changes of such a cut-off parameter around the chosen value. The limit value of income is equal to 34.4 kPLN.

4. Analysis and results

In the first step of the analysis we investigated an impact of q parameter on an accuracy of the results and tree complexity. We were changing q in the range from 0 to 50 with a step of 0.25. Increasing q beyond 50 did not change results. The algorithm was trained on the data sample which consisted of 50% randomly selected objects from the population. Trees for all values of q were tested on the whole data set. After a test, the trees were pruned by excluding leaves for which the exclusion did not cause a decrease of accuracy (percentage of correctly classified objects). During the second step of the analysis we selected the best tree based on the defined measures. We presented a structure and discussed the results provided by the chosen tree.

4.1. Q-entropy for classification

In order to quantitatively compare trees obtained for different values of q we defined a set of measures. The first three: *Acc* (*accuracy*), *Tpr* (*true positive rate*) and, *Auc* (*area under ROC curve*) are related to efficiency and effectiveness of the tree. These measures have been

TABLE I

The attributes of the households and their values.

Group	Attribute	Attribute values
1	Sex of a family head (*SEX*)	(1) male, (2) female
	Education of a family head (*EDU*)	(1) tertiary, (2) post-secondary, (3) upper secondary vocational, (4) upper secondary general, (5) basic vocational, (6) lower secondary, (7) primary, (8) no formal education
	Age of a family head (*AGE*)	16–102 (years)
	Economic group of a household (*EGRO*)	(11) employed in manual labor position, (12) employed in non-manual labor position, (2) farmer, (3) self-employed, (41) retired, (42) pensioner, (5) maintained from non-earned sources
2	Family type (*FTYPE*)	(1) marriage without children, (2–5) marriage with 1 to 4 children, (6) mother with children, (7) father with children, (8) marriage with children and other persons, (9) mother with children and other persons, (10) father with children and other persons, (11) other persons with children, (12) singles, (13) others
	Number of persons in a household (*NPER*)	1–15
	Number of children (*NCHIL*)	0–9
	Number of earners (*NEAR*)	1–10
3	Place of residence (*PRES*)	(1) town $\geq$ 500, (2) town 200–499, (3) town 100–199, (4) town 20–99, (5) town $<$ 20, (6) village (thousands of residents)
	Voivodeship (*VOI*)	(2) dolnosl., (4) kuj.-pom., (6) lubel., (8) lubus., (10) lodzk., (12) malopol., (14) mazow., (16) opolsk., (18) podkarp., (20) podlas., (22) pomor., (24) slask., (26) swietok., (28) warm.-mazur., (30) wielkopol., (32) zachodniopom.

often used in economics for evaluation of classification models in the context of e.g. credit scoring [11], income and poverty determinants [21]. The *Auc* is regarded as the additional tree quality criterion, which can be used to evaluate obtained trees [22, 23]. The measures are explained in the next paragraph. The next measure *Lev* expresses a complexity of the tree as the number of its leaves. This measure favors small trees which usually lead to simple and general rules, thus having an advantage over other models. A good tree shall be characterized by the high accuracy and *Auc* area as well as the relatively small number of leaves. In other words we would like to obtain small but efficient structures.

We deal with a problem of binary classification, in which the model yields to two results: positive and negative. There are four possible outcomes, as shown in Table II.

TABLE II

Confusion matrix for binary classification.

Predicted	Observed	
	Positives	Negatives
Positives	True Positives (TP)	False positives (FP)
Negatives	False Negatives (FN)	True Negatives (TN)

We define classification accuracy as a number of correctly identified objects divided by a number of all objects: $Acc = (\mathrm{TP} + \mathrm{TN})/(\mathrm{TP} + \mathrm{FN} + \mathrm{TN} + \mathrm{FP})$. In order to construct *Auc* measure we need to define two more indicators: $Tpr = \mathrm{TP}/(\mathrm{TP} + \mathrm{FN})$ and $Fpr = \mathrm{FP}/(\mathrm{FP} + \mathrm{TN})$ as well as a ROC curve. As mentioned earlier, each tree's node and leaf has a class assigned based on the share of HIGH objects. If the share exceeds the decision threshold, usually set to 0.5, a node or leaf gets a class HIGH assigned, otherwise class REST. Defined indicators can be calculated for various values of the decision threshold. The increase of the threshold from 0 to 1 will yield to a series of points (*Fpr*, *Tpr*) forming the curve showed in the Fig. 2. The curve is named *receiver operating characteristics, ROC* [24, 25]. Curves for the random model (random classification) and the ideal model are also presented in Fig. 2. The latter model reflects the structure of our data: 10% HIGH and 90% REST. We defined the measure *Auc* as an area under the *ROC* curve. The larger the *Auc* the model is closer to the ideal model thus the better is its performance.

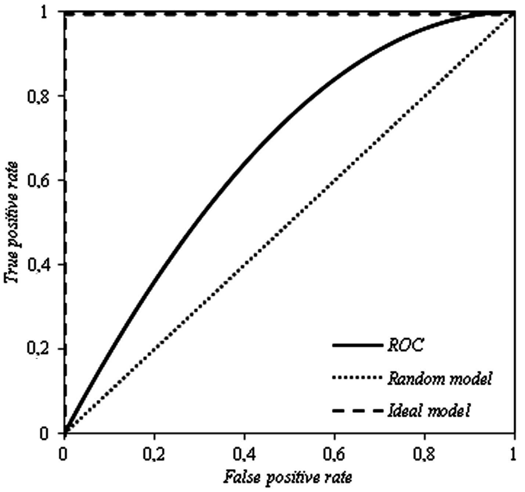

Fig. 2. ROC curve.

The measures *Acc*, *Tpr* and *Lev* were calculated for each tree. Values of measures *Acc* and *Tpr* are plotted in Fig. 3 as the function of q. Values of both measures decline rapidly at $q = 12$, the *Acc* to the level of 0.9 and *Tpr* to the value close to 0.0. This means that for $q > 12$ the percentage of correctly classified HIGH objects is close to 0% while for REST objects it's about 90%. We observe a total discrimination of the class HIGH. That's why we limit analyzed values of q to 12. Trees were also evaluated for a few additional values of $q \in \{0.05,\ 0.125,\ 0.15,\ 0.35,\ 0.65\}$. We analyzed 53 trees for $0 < q \leq 12$.

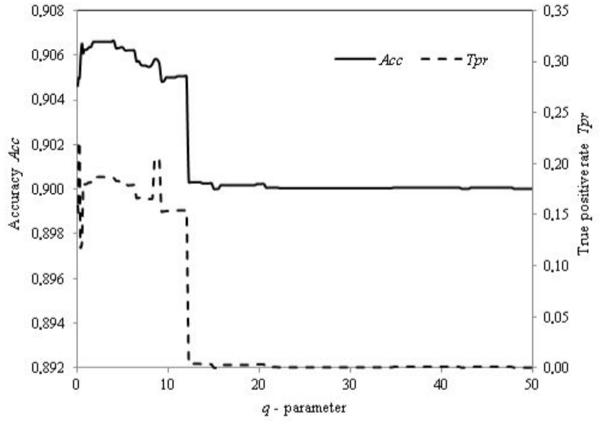

Fig. 3. The measures of the effectiveness of the trees: *Acc* and *Tpr* as a function of q.

Values of *Acc*, *Tpr*, *Auc*, *Lev* measures are shown in the Fig. 4, 5 as a function of q. The *Acc* measure has very small variability ($0.905 < Acc < 0.907$) so it is of no use during a selection of q. On the other hand a strong variance of *Tpr* is observed. We can distinguish one big local maximum for q about 0.20 and the wide maximum for q between 8.5 and 9.0. This behavior is not consistent with other measure *Auc*, which has a local maximum around $q = 1.25$. Its value is $Auc = 0.845$, about the same as for $4.25 < q < 7.00$. Each of the measures gives different value of q as the best result. The next measure taken into account is the number of leaves. Starting from $q \cong 0$, there is a strong increase of *Lev* followed by the maximum of 27 at $q = 1.5$. Then, the number of leaves decreases till $q = 9.25$ and then slightly rises.

High values of *Acc* (about 0.9) and relatively low level of *Tpr* (0.17 in average) are characteristic for classification of unbalanced data, presented in this paper. An essence of such a data is a predominance of objects of one class over objects of the other class (90% REST and 10% HIGH in our case). Trees constructed based on an unbalanced data set favor the majority class, thus yielding to high *Acc* and low *Tpr*.

4.2. Optimal q

We used a few measures to evaluate quality of obtained trees. They gave inconsistent results; each measure indicated different values of q as the best results. We decided to use three measures simultaneously. The most effective trees could be characterized by maximal *Auc* and *Tpr* and minimal *Lev*. Each of the constructed trees is described by the three numbers (*Auc*, *Tpr*, *Lev*) — the point in $\mathbf{R}^3$. An quality of the tree can be evaluated using the Euclidean distance in $\mathbf{R}^3$ between the tree and the tree pattern $(Auc_0, Tpr_0, Lev_0) = (0.846, 0.222, 6)$, where coordinates are the best observed values of the measures. The variables *Auc*, *Tpr* and *Lev* were normalized according to the formula

$$\bar{X} = \frac{X - X_{\min}}{X_{\max} - X_{\min}}. \quad (7)$$

Then the $\overline{Auc}$, $\overline{Tpr}$, $\overline{Lev} \in [0;\ 1]$ and $\overline{Auc_0}$, $\overline{Tpr_0}$, $\overline{Lev_0} = (1,\ 1,\ 0)$.

The $Dist(q) = \left((\overline{Auc}-1)^2 + (\overline{Tpr}-1)^2 + (\overline{Lev}-0)^2 \right)^{0.5}$ is presented in the Fig. 6 for all the analyzed values of q. Even the results for various measures have been inconsistent with each other the result for *Dist* is unambiguous.

One can see big variations of *Dist* for $q < 1$ while further *Dist* become more stable. It decreases with q to reach a deep local minimum. Next, there is an increase and then stabilization at a level of 0.8. The minimal *Dist* is observed for $q = 8.75$. The optimal tree classifies correctly 90.6% of all the objects and 20.5% of HIGH objects. The *Auc* measure for the selected tree is 0.82.

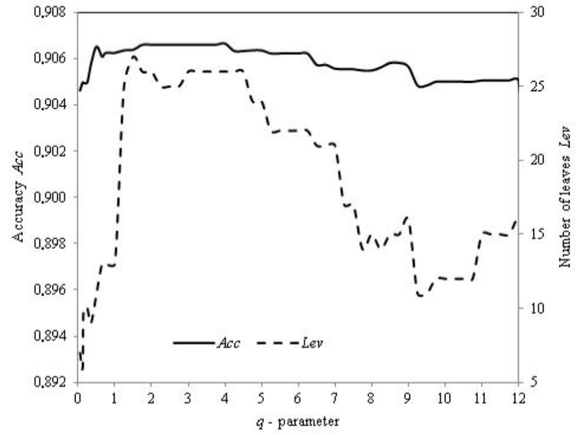

Fig. 4. The *Acc* and *Lev* measures for $0 < q \leq 12$.

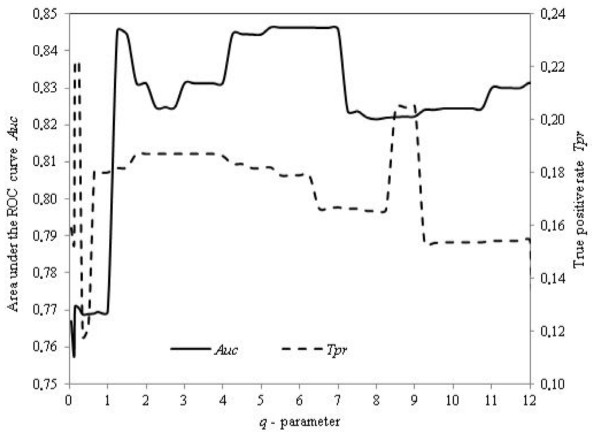

Fig. 5. The *Auc* and *Tpr* measures for $0 < q \leq 12$.

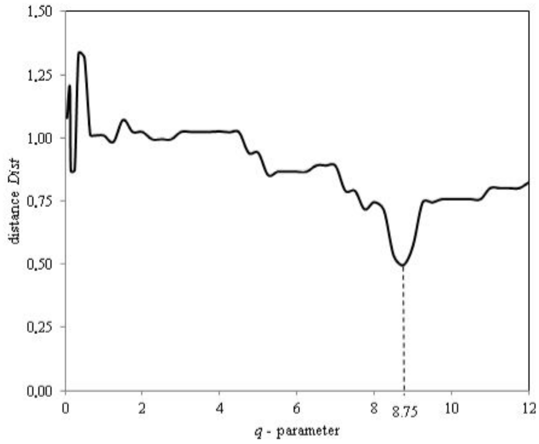

Fig. 6. *Dist* as a function of q. The minimum at $q = 8.75$ is indicated.

In order to compare results we put all the trees in Fig. 7. The optimal tree is indicated by a circle. Values of q are provided as data labels. We can distinguish two separate groups of points. The first one, for $Auc < 0.77$ is characterized by the values of $q \leq 1$. The second one, for $Auc > 0.82$ has intermediate values of Tpr and $q > 1$.

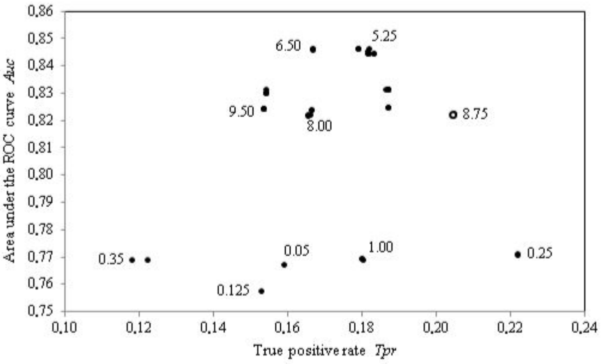

Fig. 7. *Auc* vs *Tpr* for all trees. The optimal tree is for $q = 8.75$.

4.3. Rules for high incomes

In this part of the paper we present and discus the best tree in the context of its efficiency of distinguishing of high income households. The selected classification tree for $q = 8.75$ is presented in Fig. 8. The tree has 14 leaves on 6 levels. Each node and leave has indicated: decision rule, entropy, type (HIGH or REST), and percentage of objects belonging to the majority class. There are only three leaves of type HIGH, located on levels 5 and 6 yielding to the three rules.

Rule 1 — if $EGRO = 12, 2, 3$ & $EDU = 1$ & $FTYPE = 1, 2, 3, 4, 5, 6, 7, 12$ & $VOI = 14, 22$ & $NEAR = 2$ then the outcome is a group in which a proportion of households with high incomes is 62.8 percent.

Rule 2 — if $EGRO = 12, 2, 3$ & $EDU \neq 1$ & $NEAR = 1$ & $FTYPE = 1, 2, 3, 4, 5, 8, 11$ & $VOI \neq 14, 16, 32$ & $EDU = 2, 3$ then the outcome is a group in which a proportion of households with high incomes is 51.8 percent.

Rule 3 — if $EGRO = 12, 2, 3$ & $EDU \neq 1$ & $NEAR = 1$ & $FTYPE = 1, 2, 3, 4, 5, 8, 11$ & $VOI = 14, 16, 32$ & $NPER \neq 3$ then the outcome is a group in which a proportion of households with high incomes is 93.1 percent.

The following attributes were used during building rules for those leaves: (1) *EGRO — Economic Group of Household*; (2) *EDU — Education*; (3) *FTYPE — family type*; (4) *NEAR — Number of earners*; (5) *VOI — Voivodeship*; (5) *NPER — Numer of persons in the household*. The *EGRO* as well as *EDU* were the most important attribute for all the rules. Moreover, all the rules distinguish the same values of *EGRO* which corresponds to the biggest household income coming from: employment in non-manual labor position, work as a farmer, self-employment. Another issue is to evaluate ranking of attributes based on their importance. In order to judge an attribute importance we take into account a level of the tree at which an attribute has been used for splitting. A review of obtained trees shows that some of the attributes are more important than others. They appear on the highest levels of the trees (1–3) irrespective of the value of q. The results are presented in Table III as the sum over $0 < q \leq 12$.

TABLE III

Variable importance.

Variable	Tree level					
	1	2	3	4	5	6
Voivodeship (*VOI*)			29	34	47	41
Place of residence (*PRES*)		2		23	51	27
Number of persons in a household (*NPER*)			2	1		28
Number of children (*NCHIL*)						
Number of earners (*NEAR*)			50	57	22	4
Age of a family head (*AGE*)		4	2	5	64	44
Sex of a family head (*SEX*)			1		3	
Family type (*FTYPE*)		56	17	45	40	19
Education of a family head (*EDU*)	11	28	48	16	1	3
Economic group of a household (*EGRO*)	42	10	11	10		

We could observe that for the first level (first split) the only two variables were the most important: *Economic group of a household* and *Education of a family head.* The trees' splits on the second and third level were dominated by *Family type*, *Education of a family head* and *Economic group of a household.* On the subsequent levels the following attributes become important: *Number of earners*, *Voivodeship* and *Place of residence.* The interesting finding regards features with the low predictive power which are unable to generate splits resulting in high classification accuracy. There is no influence on the classification of *Number of children.* Very small, almost negligible influence comes from the attributes: *Sex of a family head* and *Number of persons in a household.*

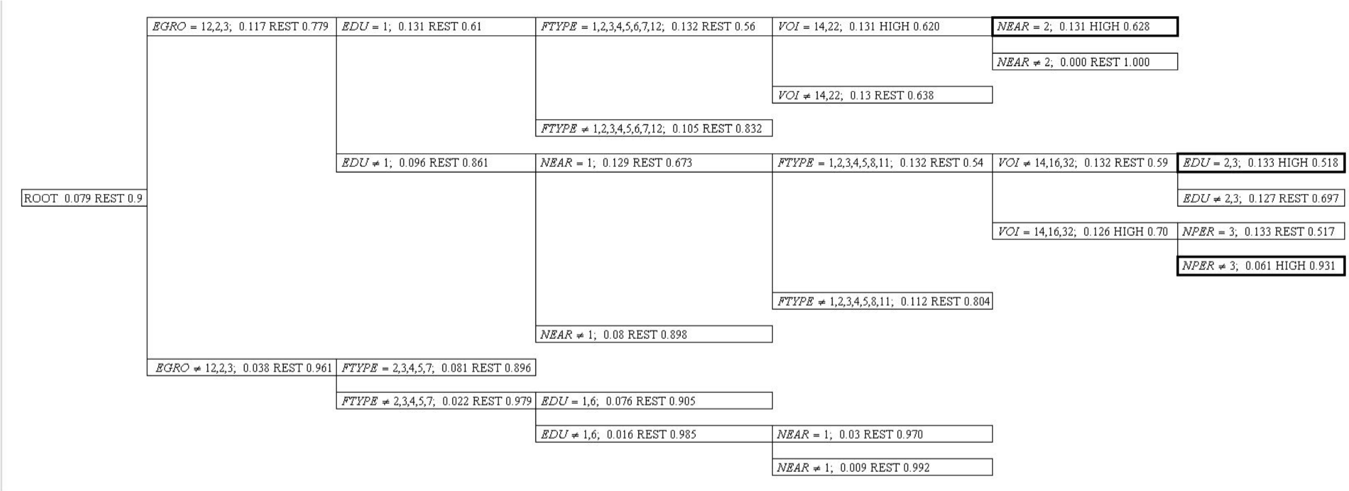
ROOT 0.079 REST 0.9
EGRO = 12,2,3; 0.117 REST 0.779
EDU = 1; 0.131 REST 0.61
FTYPE = 1,2,3,4,5,6,7,12; 0.132 REST 0.56
VOI = 14,22; 0.131 HIGH 0.620
NEAR = 2; 0.131 HIGH 0.628
NEAR ≠ 2; 0.000 REST 1.000
VOI ≠ 14,22; 0.13 REST 0.638
FTYPE ≠ 1,2,3,4,5,6,7,12; 0.105 REST 0.832
EDU ≠ 1; 0.096 REST 0.861
NEAR = 1; 0.129 REST 0.673
FTYPE = 1,2,3,4,5,8,11; 0.132 REST 0.54
VOI ≠ 14,16,32; 0.132 REST 0.59
EDU = 2,3; 0.133 HIGH 0.518
EDU ≠ 2,3; 0.127 REST 0.697
VOI = 14,16,32; 0.126 HIGH 0.70
NPER = 3; 0.133 REST 0.517
NPER ≠ 3; 0.061 HIGH 0.931
FTYPE ≠ 1,2,3,4,5,8,11; 0.112 REST 0.804
NEAR ≠ 1; 0.08 REST 0.898
EGRO ≠ 12,2,3; 0.038 REST 0.961
FTYPE = 2,3,4,5,7; 0.081 REST 0.896
FTYPE ≠ 2,3,4,5,7; 0.022 REST 0.979
EDU = 1,6; 0.076 REST 0.905
EDU ≠ 1,6; 0.016 REST 0.985
NEAR = 1; 0.03 REST 0.970
NEAR ≠ 1; 0.009 REST 0.992

Fig. 8. Decision tree for $q = 8.75$. Leaves of HIGH type are highlighted.

5. Summary and concluding remarks

We analyzed households' incomes in Poland using the entropy approach for classification. We modified a classical decision tree algorithm C 4.5 which maximizes information gain. We extended the diversity measure by incorporating parameterized Tsallis entropy thus extending classification possibilities. That allowed us to study effectiveness and complexity of the trees as a function of the entropy parameter q. The algorithm for $q = 1$ become a classical case based on the Shannon entropy. Studies were performed for $0 < q \leq 50$. It turned out that for $q > 12$ obtained trees classified almost all HIGH objects incorrectly. Limiting the parameter to $0 < q \leq 12$ we calculated measures of trees' qualities: *Acc*, *Tpr* and *Auc*. A complexity of the trees was expressed as a number of leaves *Lev*.

Taking into account classification quality and tree complexity we showed that an optimal tree exists for $q = 8.75$. The optimal tree consists of 14 leaves on 6 levels. The percentage of correctly classified objects *Acc* is equal to about 90.6% while the percentage of correctly classified HIGH objects is about 20.5%. In shall be compared to the percentage of HIGH objects in the population, which is 10%. That mean we obtained results which are more two times better that for random model. The area under the ROC curve is 0.82 while random model yields to 0.5 and best model is limited to 0.95.

For comparison, the tree based on the Shannon entropy ($q = 1$) has an $Auc = 0.77$. The percentage of correctly classified HIGH objects is 18.0%, whereas accuracy is for all values of q almost the same (≈90%). Quality classification for this tree is worse than for the best tree. On the other hand a complexity of this tree is similar to that of the optimal tree: 12 leaves on 6 levels. We observe a very similar set of attributes for $q = 1$ and $q = 8.75$.

The analysis of attributes allowed determining those characteristics of households that the most differentiate them based on belonging to HIGH and REST groups. The most important attributes are: *Economic group of a household*, *Education of a family head*, *Family type*, *Number of earners, Voivodeship, Place of residence.* On the other hand we obtained attributes with legible or no effect on classification for all values of q: *Sex of a family head*, *Number of persons in a household* and *Number of children.*

The results of the analysis are to some extent consistent with the research of [26], in respect to particular variables that discriminate the high incomes from the rest. Among the most important findings which were confirmed by our research are: (i) the education of the family head (the higher the education the higher incomes); (ii) economic group of a household (the higher incomes were observed in households of head employed in non-manual labor position, farmers and self-employed); and (iii) family type (married with/without children had higher incomes). On the other hand, the hypothesis concerning the place of the residence as an important factor discriminating the incomes wasn't confirmed by our research. According to [26], the bigger the city the higher incomes per capita were observed while the incomes in rural areas were the lowest. In undertaken study we observed that place of the residence didn't have discriminating power, furthermore in the farmers' household the observed incomes were relatively often assigned to a high incomes group. Our results are also mostly in agreement with the studies [27] performed for data 1993–2004.

Acknowledgments

This study is cofounded by the European Union from resources of the European Social Fund: Project PO KL "Information technologies: Research and their interdisciplinary applications", Agreement UDA-POKL.04.01.01-00-051/10-00.

References

[1] B.M.S. van Praag, A.J.M. Hagenaars, H. van Weern, *Rev. Income Wealth* **28**, 345 (1982).

[2] P. Maitra, F. Vahid, *J. Appl. Econom.* **21**, 999 (2006).

[3] A.A. Dragulescu, V.M. Yakovenko, *Physica A* **299**, 213 (2001).

[4] A. Banerjee, V.M. Yakovenko, T. Di Matteo, *Physica A* **370**, 54 (2006).

[5] M. Jagielski, R. Kutner, M. Pęczkowski, *Acta Phys. Pol. A* **121**, B-47 (2012).

[6] M. Jagielski, R. Kutner, *Physica A* **392**, 2130 (2013).

[7] P. Łukasiewicz, A. Orłowski, *Physica A* **344**, 146 (2004).

[8] P. Łukasiewicz, K. Karpio, A. Orłowski, *Acta Phys. Pol. A* **121**, B-82 (2012).

[9] A. Lemmens, C. Croux, *J. Marketing Res.* **43**, 276 (2006).

[10] J.L. Zhanga, W.K. Härdleb, *Comput. Stat. Data An.* **54**, 1197 (2010).

[11] M. Chrzanowska, E. Alfaro, D. Witkowska, *Expert Syst. Appl.* **36**, 6409 (2009).

[12] K. Karpio, G. Koszela, P. Łukasiewicz, A. Orłowski, *Quantitative Methods in Economics* **15** 403 (2014).

[13] J.N. Morgan, J.A. Sonquist, *J. Am. Stat. Assoc.* **58**, 415 (1963).

[14] L. Breiman, J.H. Friedman, R.A. Olshen, C.J. Stone, *Classification and Regression Trees*, Wadsworth, Belmont, CA 1984.

[15] J. Quinlan, *C 4.5: Programs for machine learning*, Morgan Kaufmann, San Mateo, CA 1993.

[16] C.E. Shannon, *Bell Syst. Tech. J.* **27**, 379, 623 (1948).

[17] C. Tsallis, *J. Stat. Phys.* **52**, 479 (1988).

[18] A. Rényi, "On measures of entropy and information", *Proceedings of the fourth Berkeley Symposium on Mathematics, Statistics and Probability 1960*, 547 (1961).

[19] T. Maszczyk, W. Duch, *Artif. Intell. Soft Comput.–ICAISC* **5097**, 643 (2008).

[20] Central Statistical Office, *Household Budget Survey in 2008*, Warsaw 2009.

[21] D. Madden, *J. Econom. Inequality* **9**, 23 (2011).

[22] A.P. Bradley, *Pattern Recogn.* **30**, 1145 (1997).

[23] J. Huang, C.X. Ling, *IEEE T. Knowl. Data En.* **17**, 299 (2005).

[24] D.J. Hand, R.J. Till, *Mach. Learn.* **45**, 171 (2001).

[25] T. Fawcett, *Mach. Learn.* **31**, 1 (2004).

[26] M. Piekut, *Polskie gospodarstwa domowe — dochody, wydatki i wyposażenie w dobra trwałego użytkowania*, Wyd. SGGW, Warszawa 2008.

[27] L. Zienkowski, Z. Żółkiewski, *Wiadomości Statystyczne* **11**, 24 (2006).