DOI: 10.15439/2015F121

Comparison of Decision Trees with Rényi and Tsallis Entropy Applied for Imbalanced Churn Dataset

Krzysztof Gajowniczek, Tomasz Ząbkowski, Arkadiusz Orłowski
Department of Informatics, Warsaw University of Life Sciences,
Nowoursynowska 159, 02-776Warsaw, Poland
Email: krzysztof_gajowniczek@sggw.pl, tomasz_zabkowski@sggw.pl, arkadiusz_orlowski@sggw.pl

Abstract—**Two algorithms for building classification trees, based on Tsallis and Rényi entropy, are proposed and applied to customer churn problem. The dataset for modeling represents highly unbalanced proportion of two classes, which is often found in real world applications, and may cause negative effects on classification performance of the algorithms. The quality measures for obtained trees are compared for different values of α parameter.**

I. Introduction

DECISION trees are powerful and very popular tools for different classification tasks [1]-[3]. The attractiveness of this technique is due to the fact that they create rules that can be easily interpreted. Decision trees use some statistical property called information gain to measure the classification power of the input attributes on classification problem as the difference between the entropy before and after a decision. Entropy computation is used to generate simple decision trees, in terms of the structure, with effective classification, since tree size reduction depends on the attribute selection. For this purpose, usually Shannon entropy is used, but other entropy formulas, such as Rényi [4] and Tsallis [5] entropy, can also be applied. Here, a comparative study based on Rényi and Tsallis entropy is described taking into account the issue of imbalance in the class distribution. We used data from telecommunication industry to predict loss of customers to competitors what is known as customer churn. In this dynamical and liberal market customers can choose among cellular service providers and actively migrate from one service provider to another. This problem is especially interesting due to the fact that the portion of churning customers in business practice is low, between 1% and 5%, depending on the country and type of the telecommunication service.

The comparison of the trees is carried out by taking into account different values of α parameter and set of the following measures: classification accuracy, area under the ROC curve, lift, and number of leaves in a tree as complexity measure.

In the next section properties of Rényi and Tsallis entropies are described. The data used in this study are described in the third section. The empirical analysis and comparison of the entropies is shown in fourth section. This type of analysis is especially interesting for decision trees because of the high dimensionality of telecommunication data. Conclusions are given in the last section.

II. Theoretical Framework

In this paper we assume that observations may belong to two given classes and for the classification we use a modified algorithm similar to C4.5 [6] to construct a binary tree in R environment [7].

As a general measure of diversity of objects, a Shannon entropy is often used which is defined as [8]:

$$H_s = -\sum_{i=1}^{n} p_i \log p_i, \qquad (1)$$

where p_i is the probability of occurrence of an event x_i being an element of the event X that can take values $x_i,\ldots,x_n$. The value of the entropy depends on two parameters: (1) disorder (uncertainty) and is maximum when the probability p_i for every x_i is equal; (2) the value of n. Shannon entropy assumes a tradeoff between contributions from the main mass of the distribution and the tail. To control both parameters two generalizations were proposed by Rényi [4] and Tsallis [5].

The Rényi entropy is defined as:

$$H_R = \frac{1}{1-\alpha}\log\left(\sum_{i=1}^{n} p_i^{\alpha}\right), \qquad (2)$$

where parameter α is used to adjust the measure depending on the shape of probability distributions.

The Tsallis entropy is defined as:

$$H_R = \frac{1}{\alpha-1}\left(1-\sum_{i=1}^{n} p_i^{\alpha}\right) \qquad (3)$$

With Shannon entropy, events with high or low probability have equal weights in the entropy computation. However, using Tsallis entropy, for $\alpha > 1$, events with high probability contribute more than low probabilities for the entropy value [9]. Therefore, the higher is the value of α, the higher is the contribution of high probability events in the final result. Furthermore, increasing α parameter $(\alpha \to \infty)$ makes the Rényi entropy determined by events

978-83-60810-66-8/$25.00©2015, IEEE

with higher probabilities, and lower values of α coefficient $(\alpha \to 0)$ weigh the events more equally, no matter of their probabilities.

The Tsallis and Rényi entropies were successfully applied to many diverse practical problems, showing their high usefulness for accurate classification. For instance, in [10] the authors applied both entropies for variable selection in computer networks intrusion detection, analyzing models detection capabilities while providing a set of attributes coming from the network traffic. Their results showed that selecting attributes based on Rényi and Tsallis entropies can achieve better results as compared to Shannon entropy.

Modified C4.5 decision trees based on Tsallis and Rényi entropies have been tested on several high-dimensional microarray datasets in [11]. The results showed that use of non-standard entropies may be highly recommended for this kind of data.

In [12] the authors addressed the question whether the Rényi entropy is equally suit-able to describe systems with q-exponential behavior, where the use of the Tsallis entropy is relevant. The study confirmed that in this case Tsallis entropy is a more suitable choice than Rényi entropy.

Some other studies considered image segmentation based on Tsallis and Rényi-entropies [13]. Their conclusion was that entropic segmentation can give good results but is highly related to an appropriate choice of the entropic index α.

III. The churn dataset

Customer churn is a term used in the telecommunication industry to describe the customer movement from one provider to another, and the churn management strategy is a process aimed to retain profitable customers [14]. Every year telecommunication industry suffers from a substantial loss of valuable customers to competitors. In this liberal market customers can migrate between telecommunication operators freely. The motivation for churn research is based on the fact that it costs more to recruit new customers than to retain existing ones, especially those high profitable customers. The other motivation is the fact that the average churn at cellular providers is about 25% per year in Europe, according to [15], what means that one fourth of the customers' base is lost each year.

In order to check the performance of the proposed entropies, we conducted the simulations based on the data collection known as "Cell2Cell: The Churn Game" [16] derived from the Center of Customer Relationship Management at Duke University, USA. The data constitute a representative slice of the entire customer database, belonging to an anonymous company operating in the sector of mobile telephony in the United States.

The data contains 71047 observations, wherein each observation corresponds to the individual customer. For each observation 78 variables are assigned, of which 75 potential explanatory variables are used for models construction. All explanatory variables are derived from the same time period, except the binary dependent variable (the values 0 and 1) labeled as "churn", which has been observed in the period from 31 to 60 days later than the other variables. In the collection there is an additional variable "calibrat" to identify the learning sample and test sample, comprising 40000 and 31047 observations, respectively. Learning sample contains 20000 cases classified as churners (leavers) and 20000 cases classified as non-churners. In the test sample, which is used to check the quality of the constructed model, there is only 1.96% of customers who quit. Such a small percentage of the modeled class can be often found in the business practice.

IV. Analysis and results

A. Accuracy measures

To compare the trees obtained for different values of α we define a set of three measures. These are: (1) AUC (area under the ROC curve), (2) Lift and, (3) Lv (number of leaves in a tree). The first two measures are related to efficiency and effectiveness of the tree and they have been often used for evaluation of classification models in the context of e.g. credit scoring [17], income and poverty determinants [18] or customer insolvency and churn [19]. The last measure Lv expresses a complexity of the tree as the number of its leaves. In this study we will favor small trees which usually lead to simple and general rules, thus having an advantage over other models. Therefore, a good tree will be characterized by the high accuracy of AUC and lift as well as the relatively small number of leaves. In other words we would like to obtain small but efficient structures for churn classification.

Since we deal with a problem of binary classification, the model yields two results: positive and negative. There are four possible outcomes, as shown in Table 1.

In order to construct AUC measure we need to define two indicators: $Tpr = TP / (\mathrm{TP+FP})$, $Fpr = FP / (\mathrm{FP}+T\,\mathrm{P})$ as well as a ROC curve. As mentioned earlier, each tree's node and leaf has a class assigned based on the share of churn classes. If the share exceeds the decision threshold, usually set to 0.5, a node or a leaf gets a class churn=1 assigned, otherwise class churn=0.

Table 1.
Confusion matrix for binary classification

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

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 with Tpr on horizontal axis and Fpr on vertical axis. The curve is named receiver operating characteristics, ROC [20], [21]. The AUC measure is an area under the ROC curve which can be calculated using trapezoidal rule. Theoretically $AUC \in [0;1]$ and the larger the AUC the closer is the model to the ideal one and the better is its performance.

The lift measure is dictated by the economic considerations, because the telecom operator does not direct the retention campaign to a wide customer base, but focuses on a small percentage of approximately 1-2% of the customer database on a monthly basis, characterized by the highest probability of resignation. For instance, having the total number of customers of approximately 10 million, a group of 1% of customers is equal to 100 thousand customers per month, which would receive the retention offer.

The required input for lift calculation is a validation dataset that has been “scored" by assigning the estimated churn probability to each case. Next, the churn probabilities are sorted in descending order and for a given customers percentage, the measure is calculated in the following manner (for the first percentile) [22]:

$$Lift_{0.01} = \frac{TP_{0.01}}{TP} \tag{4}$$

The lift measure shows how much more likely we are to receive positive responses (detecting churn customers) in comparison to a random sample of customers.

B. Experiments

Rényi and Tsallis entropy were compared to each other using the modified C4.5 algorithm for decision tree construction which has been applied to churn dataset. The modification of the algorithm concerned mainly the pruning part. The listing *Generate_decision_tree* presents the tree growing algorithm.

The algorithm is recursively called so that it works from the bottom of the tree upward, removing or replacing branches to minimize the predicted error on the validation dataset.

In order to obtain the optimal split while growing the tree (see part of the pseudo-code above) the gain ratio should be calculated. The listing *Prune* outlines the pruning process.

The algorithm is recursively called so that it works from the bottom of the tree up-ward, removing or replacing branches to minimize the predicted error on the validation dataset.

The decision trees were trained on training samples which reflected two designs: (1) learning on the balanced dataset (equal proportion of churn and non-churn classes); (2) learning on the imbalanced dataset with the churn rate of 1.96%. Both designs were then checked on the validation sample in which the churn rate was equal to 1.96%, as observed in real population. We considered α starting from 0.5 to 10 by 0.5.

The results obtained on the validation datasets are collected in Tables 2-3. The best results and corresponding values of α parameter differ in each case and can be summarized as follows:

i. Training the trees on the balanced dataset resulted in better classification performance;

ii. The Rényi entropy based trees trained on imbalanced dataset generated the splits only for α equal to 1 and 1.5; all the other α resulted in no split (Lv=1)

iii. In general, the Tsallis entropy based trees provided better generalization (smaller number of leaves) and the highest lift (3.612);

iv. The Rényi entropy based trees provided complex tree structures with questionable generalization abilities (although the high AUC observed).

v. The Shannon based tree trained on the balanced dataset resulted in high AUC and high lift; however the tree was very complex. The tree trained on the imbalanced dataset did not generate any splits.

Algorithm: *Generate_decision_tree*
Input: training samples **D**, list of attributes **L**, attribute_selection_method
Output: decision tree

```
Create a node N
if D has the same class C then
    return N as leaf node with class C label
if L is empty then
    return N as leaf node with class label that is the most class in D
Choose test-attribute a that has the most Gain-Ratio using attribute_selection_method
Give node N with test-attribute label
Find an optimal split that splits D into subsets Di ($i = 1,...,k$)
foreach $i = 1$ to $k$ do
    Add branch in node N to test-attribute = $a_i$
    Make partition for sample Di from samples where test-attribute = $a_i$
    if Di is empty then
        attach leaf node with the most class in D
    else attach node that generate by Generate_decision_tree(Di, attribute-list, test-attribute)
endfor
return N
```

The structure of the best tree, in terms of the lift, trained on the balanced dataset using Tsallis entropy is presented in Fig. 1. The tree has 27 leaves on 7 levels (including the root). Each node and each leave has indicated: decision rule, class (TRUE – if churn was observed and FALSE otherwise), and percentage of objects belonging to the majority class.

The first variable used for split was *EQPDYAS* (number of days of the current equipment). If the value of *EQPDYAS* was greater than 302 then the probability of churn increased, forming a group in which the percentage of churners amounted to 56.9%. On the other levels of the tree it was observed that the following variables were useful for detecting churners: *MONTHS* (months in service), *MOU* (mean monthly minutes of use), *RETCALLS* (number of calls previously made to retention team), *RECCHRGE* (mean total recurring charge), *RETACCPT* (number of

```
Algorithm: Prune
Input: node with an attached subtree, validation samples W
Output: pruned tree

        leafError = estimated leaf error on W
        if node is a leaf then
          return leaf error
        else
          subtreeError = $\sum_{N_i \in children(\text{node})} \Pr une(N_i)$
          branchError = error if replaced with most frequent branch
          if leafError is less than branchError and subtreeError then
            make this node a leaf
            error = leafError
          else if
            branchError is less than leafError and subtreeError then
              replace this node with the most frequent branch
              error = branchError
          else
            error = subtreeError
          return error
end
```

previous retention offers accepted), *PEAKVCE* (mean number of in and out peak voice calls), *DIRECTAS* (mean number of director assisted calls), *MOUREC* (mean unrounded MOU received voice calls), *CHANGEM* (% change in minutes of use), *CALLWAIT* (mean number of call waiting calls), *INCOME* (customer income), *REVENUE* (mean monthly revenue).

The final leaves contained the high proportion of churners ranging from 54.9% to 100%. Three rules lead to the leaves with 100% of churners. These were:

Rule 1 – *EQPDAYS* <= 302 & *MONTHS* <= 10 & *RECCHRGE* <= 37.8775 & *RETACCPT* > 0 & *DIRECTAS* <= 2.2275;

Rule 2 – *EQPDAYS* > 302 & *MONTHS* <= 12 & *RETCALLS* > 0 & *PEAKVCE* > 122.33 & *MOUREC* > 86.35 & *RETACCPT* <= 0;

Rule 3 – *EQPDAYS* > 302 & *MONTHS* > 12 & *MOU* <= 6 & *MOU* > 0 & *EQPDAYS* <= 375 & *CHANGEM* > -3.25.

The results presented in this paper are encouraging and provide high accuracy of classification, when compared to similar studies on this dataset. For example, the authors in [23] as an assessment of the quality of the model, chose lift in the first decile, which was equal to 2.61 for the best model. Finally, in our previous study [24] we obtained lift of 3.11 for the first percentile using C&RT tree on the same dataset, while current study delivers improved results.

TABLE 2.
RESULTS ON VALIDATION DATASET WHEN TRAINING THE TREE ON BALANCED DATASET. THE BEST RESULTS FOR EACH ACCURACY MEASURES ARE PRESENTED IN BOLD

	Tsallis			Rényi		
Alpha	***AUC***	***Lift***	***Lv***	***AUC***	***Lift***	***Lv***
0,5	61,32	2,463	25	59,66	1,313	25
1	61,55	2,627	33	60,83	1,642	29
1,5	61,95	1,806	31	61,23	1,642	30
2	60,62	1,313	34	61,13	**3,284**	39
2,5	60,86	1,313	35	61,30	2,791	38
3	61,32	3,284	35	62,20	2,463	42
3,5	61,97	3,119	33	61,88	3,119	45
4	61,83	3,448	35	62,73	1,642	46
4,5	61,21	2,463	28	61,34	1,149	47
5	**62,02**	2,463	31	61,29	1,806	46
5,5	61,17	2,134	34	**63,04**	0,985	41
6	60,77	3,119	30	62,05	1,149	43
6,5	61,17	**3,612**	27	63,03	1,642	41
7	58,74	2,298	19	62,53	0,985	39
7,5	61,11	3,284	26	62,68	1,313	42
8	61,12	3,448	22	62,19	1,313	42
8,5	61,27	2,791	23	62,46	1,642	38
9	60,50	2,463	20	62,34	1,477	45
9,5	61,22	2,791	22	62,49	1,477	44
10	60,84	3,119	12	59,66	1,313	25
		Shannon				
	62,98	3,248	82			

TABLE 3.
RESULTS ON VALIDATION DATASET WHEN TRAINING THE TREE ON IMBALANCED DATASET. THE BEST RESULTS FOR EACH ACCURACY MEASURES ARE PRESENTED IN BOLD

	Tsallis			Renyi		
Alpha	***Auc***	***Lift***	***Lv***	***Auc***	***Lift***	***Lv***
0,5	50,00	1,000	1	50,00	1,000	1
1	58,99	2,791	10	**60,83**	2,463	29
1,5	58,35	2,298	9	58,89	**2,955**	6
2	**59,39**	2,627	11	50,00	1,000	1
2,5	58,98	2,791	7	50,00	1,000	1
3	57,87	1,970	6	50,00	1,000	1
3,5	58,10	2,791	7	50,00	1,000	1
4	58,24	**2,955**	11	50,00	1,000	1
4,5	50,00	1,000	1	50,00	1,000	1
5	58,36	1,970	11	50,00	1,000	1
5,5	57,75	2,463	7	50,00	1,000	1
6	58,44	2,627	7	50,00	1,000	1
6,5	50,00	1,000	1	50,00	1,000	1
7	56,44	2,134	8	50,00	1,000	1
7,5	50,00	1,000	1	50,00	1,000	1
8	58,12	2,791	7	50,00	1,000	1
8,5	50,00	1,000	1	50,00	1,000	1
9	50,00	1,000	1	50,00	1,000	1
9,5	50,00	1,000	1	50,00	1,000	1
10	50,00	1,000	1	50,00	1,000	1
		Shannon				
	50,00	1,000	1			

V. SUMMARY AND CONCLUDING REMARKS

In this paper, an evaluation of Rényi and Tsallis entropy applied to customer churn in telecommunication industry is performed. In particular, the modified C4.5 decision tree

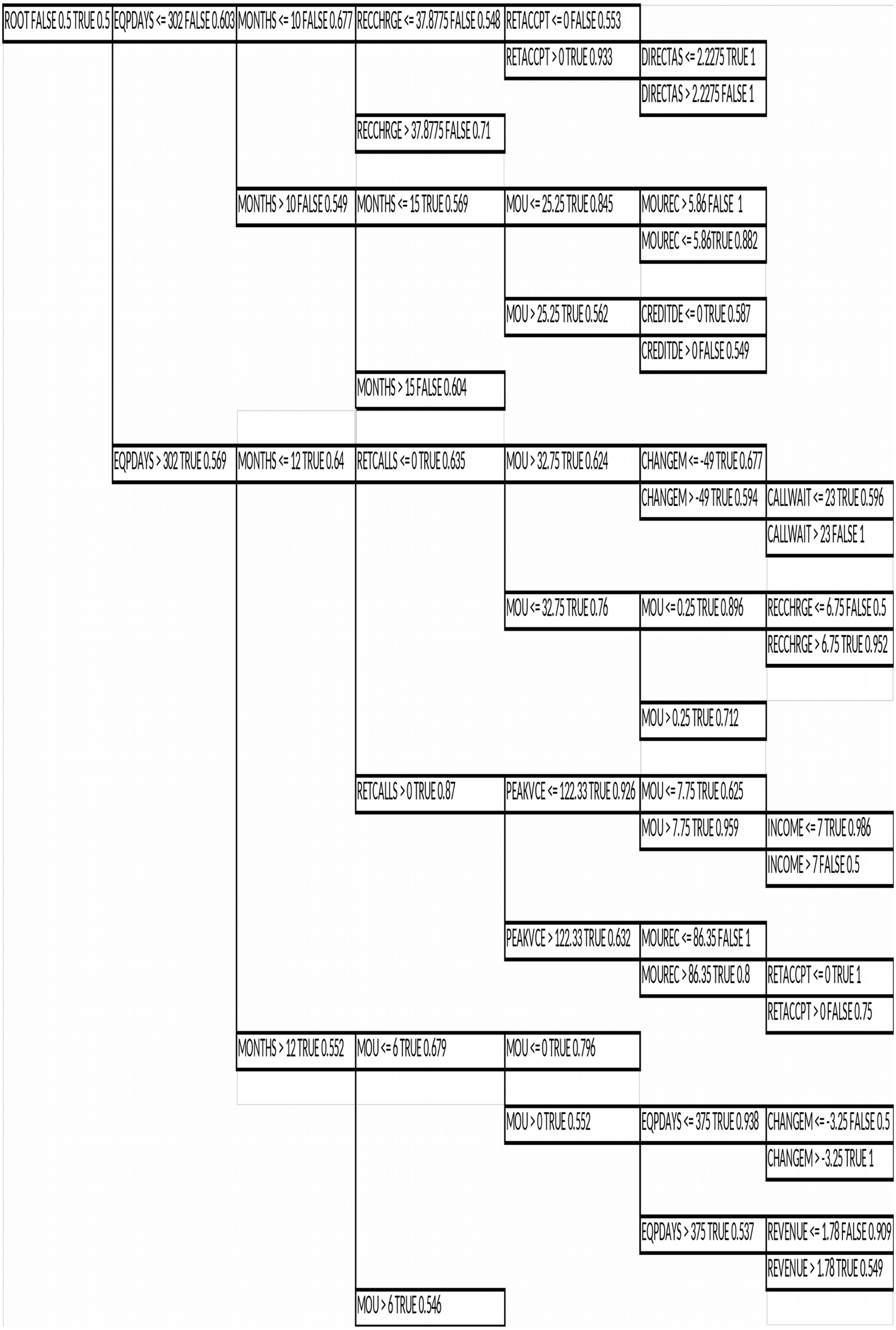

Fig. 1. Decision tree based on Tsallis entropy for α = 6.5 trained on the balanced data.

algorithm was used for classification since it can handle continuous and discrete input variables as observed in the churn dataset.

Additionally, we studied the performance of both entropies in the case of the learning dataset being balanced or imbalanced. The experimental results show that in general, Tsallis and Rényi entropies, with adequate α parameters, can lead to compact and efficient decision trees, with high accuracy measures. We observed that Tsallis entropy provided better generalization since the resulting trees were not as complex as for Rényi case. The study revealed that learning on the balanced learning dataset is beneficial for the final results. Finally, the use of Tsallis and Rényi entropies makes analysis more flexible than standard approach, e.g. Shannon entropy, since it allows for exploration of the tradeoff between the probability of different classes and the overall information gain.

Acknowledgment

The computations were performed partially employing the computational resources of the Interdisciplinary Centre for Mathematical and Computational Modeling at the Warsaw University (Computational Grant No. G59-31).

References

[1] Levashenko, V., Zaitseva, E., Pancerz, K., Gomuła, J.: Fuzzy decision tree based classification of psychometric data. In: Ganzha, M., Maciaszek, L., Paprzycki, M. (eds.) Position Papers of the 2014 Federated Conference on Computer Science and Information Systems. Annals of Computer Science and Information Systems, PTI, 3, 37–41, (2014)

[2] Popescu, A., Popescu, B., Brezovan, M., Ganea, E.: Image semantic annotation using fuzzy decision trees. In Computer Science and Information Systems (FedCSIS), IEEE, 597–601 (2013)

[3] Levashenko, V., Zaitseva, E.: Fuzzy decision trees in medical decision making support system. In Computer Science and Information Systems (FedCSIS), IEEE, 213–219 (2012)

[4] Rényi, A.: On measures of entropy and information. Proc. of the 4th Berkeley Symposium on Math. Statistics and Prob., University of California Press, Berkeley, 547–561 (1961)

[5] Tsallis, C.: Possible generalization of Boltzmann-Gibbs statistics. Journal of Statistical Physics 52(1-2), 479–487 (1988)

[6] Quinlan, J.: C 4.5: Programs for machine learning. Morgan Kaufmann, San Mateo, CA (1993)

[7] Team, R. Core. R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria, 2012. (2014)

[8] Shannon, C.E.: A Mathematical Theory of Communication. The Bell System Technical Journal 27, 379–423, 623–656 (1948)

[9] Gajowniczek, K., Karpio, K., Łukasiewicz, P., Orłowski, A., Ząbkowski, T.: Q-entropy approach to selecting high income households, Acta Physica Polonica A, 127(3A), 38–44 (2015)

[10] Lima, C.F.L., Assis de, F. M., Souza de, C.P.: A Comparative Study of Use of Shannon, Rényi and Tsallis Entropy for Attribute Selecting in Network Intrusion Detection. H. Yin et al. (Eds.) IDEAL 2012, Lecture Notes in Computer Science 7435, 492–501 (2012)

[11] Maszczyk, T., Duch, W.: Comparison of Shannon, Renyi and Tsallis Entropy Used in Decision Trees. Rutkowski et al. (Eds.): ICAISC 2008, Lecture Notes in Computer Science 5097, 643–651 (2008)

[12] Johal, R.S., Tirnakli, U.: Tsallis versus Renyi entropic form for systems with q-exponential behaviour: the case of dissipative maps, Physica A 331, 487–496 (2004)

[13] Li, Y., Fan, X., Li, G.: Image segmentation based on Tsallis-entropy and Renyi-entropy and their comparison. IEEE International Conference on Industrial Informatics, 943–948 (2006)

[14] Berson, A., Smith, S., & Thearling, K.: Building data mining applications for CRM. New York, NY: McGraw-Hill (2002)

[15] Chang, Y.T.: Applying data mining to telecom churn management. International Journal of Reviews in Computing 1(10), 67–77 (2009)

[16] Neslin, S.: Cell2Cell: The churn game. Cell2Cell Case Notes. Hanover, NH: Tuck School of Business, Dartmoth College (2002)

[17] Chrzanowska, M., Alfaro. E., Witkowska, D.: The Individual Borrowers Recognition: Single and Ensemble Trees. Expert Systems with Applications 36(3), 6409- 6414 (2009)

[18] Madden, D.: Health and income poverty in Ireland 2003–2006. Journal of Economic Inequality 9, 23–33 (2011)

[19] Ząbkowski, T., Szczesny, W.: Insolvency modeling in the cellular telecommunication industry. Expert Systems with Applications 39, 879-6886 (2012)

[20] Fawcett, T.: ROC graphs: Notes and practical considerations for researchers. Machine Learning 31, 1–38 (2004)

[21] Gajowniczek, K., Ząbkowski, T., Szupiluk, R.: Estimating the ROC curve and its significance for classification models' assessment, Quantitative Methods in Economics, 15(2), 382–391 (2014)

[22] Larose, D.T.: Discovering knowledge in data: an introduction to data mining. John Wiley & Sons (2014)

[23] Bose, I., Chen, X.: Hybrid models using unsupervised clustering for prediction of customer churn. Proceedings of the International MultiConference of Engineers and Computer Scientists I, IMECS, Hong Kong (2009)

[24] Gajowniczek, K., Ząbkowski, T.: Problems of churn modeling at cellular telecommunication (in Polish). Quantitative Methods in Economics 13(3), 65–79 (2012)