FACTA UNIVERSITATIS
Series: **Electronics and Energetics** Vol. 32, N^{o} 4, December 2019, pp. 529-538
https://doi.org/10.2298/FUEE1904529K

CLASSIFICATION OF ELECTRICITY CONSUMERS USING ARTIFICIAL NEURAL NETWORKS

Dragana Knežević, Marija Blagojević

University of Kragujevac, Faculty of Technical Sciences Čačak, Serbia

Abstract. *This paper explains the process of using neural networks, as one of numerous data mining techniques, for the classification of electricity consumers. The processed data comprised more than a million recordings of electricity consumption for 21,643 consumers over the period of four years and eight months. Using a data subset (70% of the entire dataset), the network was trained for the classification of consumers according to the type of the electric meter they possess (single-rate or dual-rate) and the zone they live in (city or village). The network input data in both cases included: consumer code, reading period from-to, current and previous meter reading for both low and high tariff, dual and single rate tariff consumption for that period and their total amount, as independent variables, whereas the network output comprised dependent variable classes (zone or type of electric meter). The results show that a network created in this way can be trained so well that it achieves high precision when evaluated using the test dataset. Using the available recordings about electricity consumption, the type of the electric meter consumers possess and the zone they live in can be predicted with the accuracy of 77% and 82%, respectively. These findings can provide the basis for further research using other data mining techniques.*

Key words: *data mining, neural network, classification, prediction, electricity, R programming.*

1. INTRODUCTION

Data mining has emerged as a result of the attempts to find more effective ways of dealing with ever-growing amounts of stored data. The appropriate use of data can be highly beneficial, and data mining is primarily aimed at discovering patterns among seemingly completely unrelated data. Depending on the problem that should be solved, the volume of available data, and the format in which results should be reported, one of numerous data mining techniques can be selected.

Data mining is commonly considered a multidisciplinary field [1], which has been proven by its applications in various areas. It is used for research purposes in economics, for testing security systems and discovering elements that affect their performance [2], for

Received January 17, 2019; received in revised form June 3, 2019

Corresponding author: Dragana Knežević
Faculty of Technical Sciences Čačak, Svetog Save 65, 32000 Čačak, Serbia
(E-mail: dknezevic28@gmail.com)

education-related purposes [3], for increasing IT efficiency and solving problems of storing huge sets of data, as well as the problems of transferring, processing and analyzing data using Internet of Things [4], and even for making predictions about the risk of falling off a bike [5], or about the result of a football match [6]. Furthermore, data mining techniques can be used to analyze electricity consumption in order to design more efficient electric power distribution systems that would meet specific consumers' needs [1], or to compare the electricity consumption during different seasons, national holidays, etc. [7]. Moreover, these techniques can be used for consumer clustering based on the amount of consumed electricity, taking into account weather conditions and other factors [8], as well as for short-term and long-term planning and predictions of electricity demand, but also for predictions of potential electricity theft and misuse [9, 10 , 11, 12]. The authors of [13], for instance, developed a model for the detection of two types of illegal electricity consumption (entire or partial) using the combination of classification methods and the Levenberg-Marquardt method in smart grid. Drawing upon the knowledge bases providing information on the existing medium voltage electricity consumers, the authors of [14] endeavored to identify typical characteristics and load profiles of consumers in order to ensure successful predictions and classification of new consumers, whereas the aim of [15] was to perform the classification and categorization of the existing consumers based on the characteristics of the electricity consumption.

This paper focuses on the classification of electricity consumers based on two different criteria: the type of the electric meter they possess and the zone they live in. An extremely huge dataset was used for the purpose of the research, exceeding a million recordings, in order to ensure higher precision of the obtained results. Such results may serve as a basis and inspiration for future research.

One of the greatest data mining challenges is the selection of the appropriate algorithm. Unlike other fields (e.g. statistics), data mining allows the simultaneous use of several different techniques, and the applied algorithms are tested using a data subset (commonly referred to as the test dataset) before selecting the one that yields the most reliable results.

In this paper, the emphasis is placed on solving the problem of classification using neural networks as a type of supervised learning. The aim of the research is to find out whether neural networks can be used for the classification of electricity consumers based on several different criteria. The very fact that the given classes might not be clearly partitioned, i.e. the fact that they might be intertwined regarding the type of the meters and the zone where they are installed, makes the research all the more interesting. It also makes it more complicated for the algorithm itself to spot and make the differences.

2. Methodology

The target dataset included the information about electricity consumers on the territory of the City of Užice during the period of four years and eight months, i.e. from January 2014 to August 2018.

The observed geographical area, i.e. the consumers on the territory under the jurisdiction of the electricity distribution company of the City of Užice – ED Užice, can be classified into several basic categories. One classification may be performed on the grounds of the place where the meters are installed, i.e. whether they are in urban or rural areas, which is an interesting classification criterion. On the other hand, there are two types of electric meters, single-rate and dual-rate ones, and the above-mentioned zones usually differ with respect to these types, i.e. most single-rate meters are installed in rural areas, though this is not a rule, not universally true. Therefore, it might be interesting to perform such a classification.

The target dataset comprised 1,048,575 readings for a total of 21,643 consumers. The analyzed data included: the consumer category, the zone where consumers live, the information whether they possess a single-rate or dual-rate meter, meter readings at the beginning and end of each month for the single rate or both rates, as well as the month (the billing period), and the total electricity consumption for both rates, expressed in kWh.

This dataset had a special target variable, which became its class attribute. Due to the fact that classes were determined by the user, this type of learning is called supervised learning [16], which implies that the training dataset is used to train a neural network, which is subsequently validated using the test dataset. The data were processed and the neural network created using the R program language, i.e. in the R Studio. After importing the dataset into the R Studio working environment, and before activating the neural network algorithm, it was necessary to perform the data preprocessing. As the used dataset contained different types of data (numeric data, strings, dates, etc.), the process of data normalization was performed and all the data transformed into numeric values belonging to the 0-1 range, as shown in Figures 2 and 3. Preprocessing also implied the differentiation between relevant data and those insignificant for the desired analysis, and it was followed by the classification into two groups, the training and test datasets.

For the training of the neural network, 70% of the data (734,003 precisely) were selected using the accidental sampling method, whereas the remaining data (314,572) were used for the purpose of testing the model (the test dataset). After the removal of irrelevant and redundant attributes, and the selection of the classification variable, the neural network algorithm was activated. The input data fed into the activation function included: the dependent variable, independent variables, the target set (i.e. the data subset used to train the network), the selected algorithm, the number of neural network training repetitions, the decision whether the output would be printed and how, the threshold value, and the number of hidden layers, if any. While training a neural network, the results were printed and the network diagram, as shown in Figure 1, was created, showing the input and output values, as well as hidden layers. Moreover, the values of attribute weights in a layer (layers) can be seen.

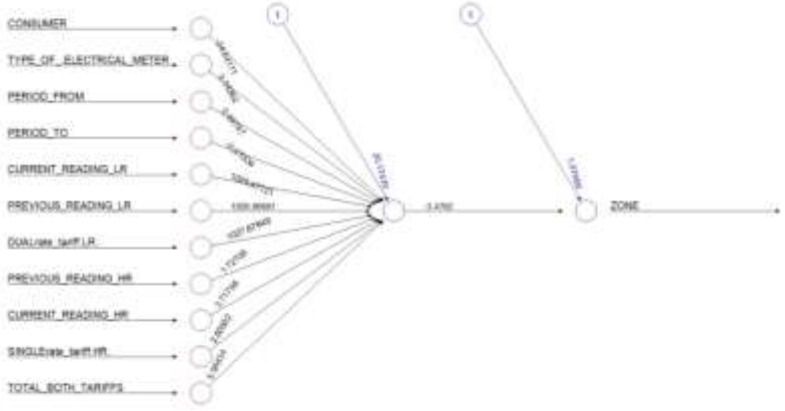

Fig. 1 Diagram of trained neural network

All the data ‘learnt’ in the previous process (with the exception of the classification variable) were evaluated using the test dataset. A table containing the predicted and actual values was created, showing the network prediction accuracy. Finally, the confusion matrix was used to summarize the values of all correct and incorrect predictions [17, 18]. A typical confusion matrix is shown in Table 1.

Table 1 Confusion matrix [17]

	Predicted value	
Actual value	Classified negative	Classified positive
Actual negative	TN	FP
Actual positive	FN	TP

TP – true positive: the model correctly predicts the positive class (we predicted ‘yes’, and it is ‘yes’),

TN – true negative: the model correctly predicts the negative class (we predicted ‘no’, and it is ‘no’),

FP – false positive: the model incorrectly predicts the positive class (we predicted ‘yes’, but it is ‘no’),

FN – false negative: the model incorrectly predicts the negative class (we predicted ‘no’, but it is ‘yes’) [16, 14].

Given a confusion matrix, other measures such as accuracy, precision (p) and recall/sensitivity (r) can be calculated using equations 1, 2 and 3, respectively.

$$Accuracy = \frac{Number\ of\ correctly\ classified\ units}{Total\ number\ of\ units\ in\ test\ dataset} = \frac{TP+TN}{P+N} \tag{1}$$

$$p = \frac{TP}{TP+FP} \tag{2}$$

$$r = \frac{TP}{TP+FN} \tag{3}$$

Although, theoretically speaking, there is no correlation between precision and recall, in practice, a high level of precision is almost always achieved at the expense of recall, and the maximum recall is achieved at the expense of precision. Which of these two measures is more important mostly depends on the nature of the application. Very often, when a single metric is needed for the comparison of different classifiers, the F-score (also known as F_1-score) is used (4) [17]:

$$\mathrm{F_1} = \frac{2\mathrm{pr}}{\mathrm{p+r}}. \tag{4}$$

The F score is the harmonic mean of precision (p) and recall (r). The harmonic mean of a pair of numbers strongly tends towards the lower one. Therefore, in order to get high F_1-score, both p and r values must be high. There is another measure, known as the precision and recall breakeven point. The breakeven point is a point at which precision equals recall (p=r). It implies that different cases can be classified based on their probability of being positive. If this point cannot be found, the interpolation must be performed, using an interpolation method [17].

3. Results and Discussion

The paper presents the results of the analysis of two different examples of prediction. The research was carried out using the same dataset, but different dependent (class) variables. The aim of the first analysis was to build a neural network that would predict whether electricity consumers possess a single-tariff or dual-tariff electric meter, whereas the aim of the second one was to predict in which zone the consumers live.

3.1. Classification of consumers according to type of meter they possess

In order to solve this problem, two classes of the dependent variable - '*TYPE_OF_ELECTRIC_METER*' were defined. A single-tariff meter was labelled '0', and a dual-tariff meter was labelled '1'. The normalization of data was done at the beginning of the research. Table 2 shows the data subset before, and Table 3 shows the same subset after the normalization process.

Table 2 Data subset overview before normalization

CONSUMER	ZONE	PERIOD_FROM	PERIOD_TO	TYPE_OF_ ELECTRICAL_METER	CURRENT_READING_LR	PREVIOUS_READING_LR	DUALrate_tariff(LR)	PREVIOUS_READING_HR	CURRENT_READING_HR	SINGLErate_tariff(HR)	TOTAL_BOTH_TARIFFS
POTROSAC_1	1	01.01.2014	31.01.2014	0	1	1	0	64567	65014	447	447
POTROSAC_1	1	01.02.2014	28.02.2014	0	1	1	0	65014	65418	404	404
POTROSAC_1	1	01.03.2014	31.03.2014	0	1	1	0	65418	65800	382	382
POTROSAC_1	1	01.04.2014	30.04.2014	0	1	1	0	65800	66199	399	399
POTROSAC_1	1	01.05.2014	31.05.2014	0	1	1	0	66199	66634	435	435
POTROSAC_1	1	01.06.2014	30.06.2014	0	1	1	0	66634	67025	391	391
POTROSAC_1	1	01.07.2014	31.07.2014	0	1	1	0	67025	67457	432	432
POTROSAC_1	1	01.08.2014	31.08.2014	0	1	1	0	67457	67823	366	366
POTROSAC_1	1	01.09.2014	30.09.2014	0	1	1	0	67823	68279	456	456
POTROSAC_1	1	01.10.2014	31.10.2014	0	1	1	0	68279	68734	455	455
POTROSAC_1	1	01.11.2014	30.11.2014	0	1	1	0	68734	69108	374	374
POTROSAC_1	1	01.12.2014	31.12.2014	0	1	1	0	69108	69565	457	457
POTROSAC_1	1	01.01.2015	31.01.2015	0	1	1	0	69565	70055	490	490
POTROSAC_1	1	01.02.2015	28.02.2015	0	1	1	0	70055	70422	367	367

In this example, the independent variables included: *CONSUMER, ZONE, PERIOD_FROM, PERIOD_TO, CURRENT_READING_LR, PREVIOUS_READING_LR, DUALrate_tariff.LR, PREVIOUS_READING_HR, CURRENT_READING_HR, SINGLErate_tariff.HR, TOTAL_BOTH_TARIFFS.* A little more than 70% of the data were used as the training dataset, and the remaining percentage served as the test dataset.

Table 3 Data subset overview after normalization

CONSUMER	ZONE	PERIOD_FROM	PERIOD_TO	TYPE_OF_ ELECTRICAL_METER	CURRENT_READING_LR	PREVIOUS_READING_LR	DUALrate_tariff(LR)	PREVIOUS_READING_HR	CURRENT_READING_HR	SINGLErate_tariff(HR)	TOTAL_BOTH_TARIFFS
1	0	0	0	0.418182	0	0	0	0	0.14096	0.141649	0.014261
2	0	0	0.090909	0	0	0	0	0	0.141936	0.142529	0.012889
3	0	0	0.181818	0.509091	0	0	0	0	0.142818	0.143362	0.012187
4	0	0	0.272727	0.090909	0	0	0	0	0.143652	0.144231	0.012729
5	0	0	0.363636	0.6	0	0	0	0	0.144523	0.145179	0.013878
6	0	0	0.454545	0.181818	0	0	0	0	0.145473	0.146031	0.012474
7	0	0	0.545455	0.690909	0	0	0	0	0.146327	0.146972	0.013782
8	0	0	0.636364	0.781818	0	0	0	0	0.14727	0.147769	0.011677
9	0	0	0.727273	0.272727	0	0	0	0	0.148069	0.148763	0.014548
10	0	0	0.8	0.872727	0	0	0	0	0.149064	0.149754	0.014516
11	0	0	0.872727	0.345455	0	0	0	0	0.150058	0.150569	0.011932
12	0	0	0.945455	0.945455	0	0	0	0	0.150874	0.151565	0.01458
13	0	0	0.018182	0.436364	0	0	0	0	0.151872	0.152632	0.015632
14	0	0	0.109091	0.018182	0	0	0	0	0.152941	0.153432	0.011708

After the completion of the learning process, which was performed using the training dataset, the neural network was created, as shown in Figure 2.

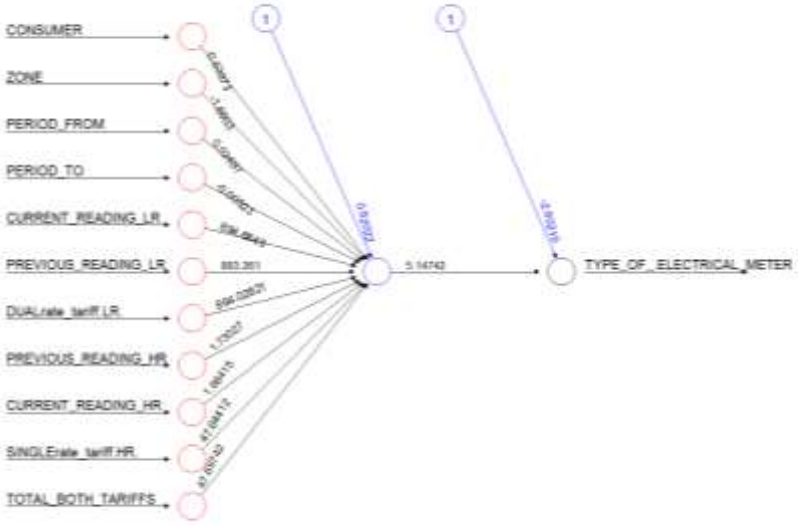

Fig. 2 Neural network diagram created using dependent variable - 'TYPE_OF_ELECTRIC_METER'

Finally, it was necessary to test the performance of the model obtained using the training dataset and determine the neural network accuracy by comparing these results with the data from the test group. First, the dependent variable was removed from the test dataset and the predictor variable was determined. Then the predicted values were compared with the actual ones, and ultimately, the confusion matrix was created, providing an overview of true, false, positive and negative results. The confusion matrix is shown in Table 4.

Table 4 Confusion Matrix

	prediction	
actual	0	1
0	77405	42398
1	27550	167220

Based on the data provided in the confusion matrix and the conclusions drawn, the accuracy of this specific model was calculated using equation 1, as well as its precision (p) using equation 2, sensitivity (r) using equation 3, and F_1-score (equation 4):

$$\text{Accuracy} = 77\%\,; \quad p = 64\%\,; \quad r = 73\%\,; \quad F_1 = 69\%.$$

For detailed information, please see Table 6.

3.2. Classification of consumers according to zone they live in

In the second example, the same dataset was used for the classification according to the zone where consumer live (urban zone – the class labeled ‘0’, rural zone – the class labeled ‘1’) in order to predict whether a consumer lives in an urban or in a rural zone. Exactly 70% of the data were used as the training, and the remaining 30% were used as the test dataset. Again, the normalization of data was done, and the neural network created, as shown in Figure 3.

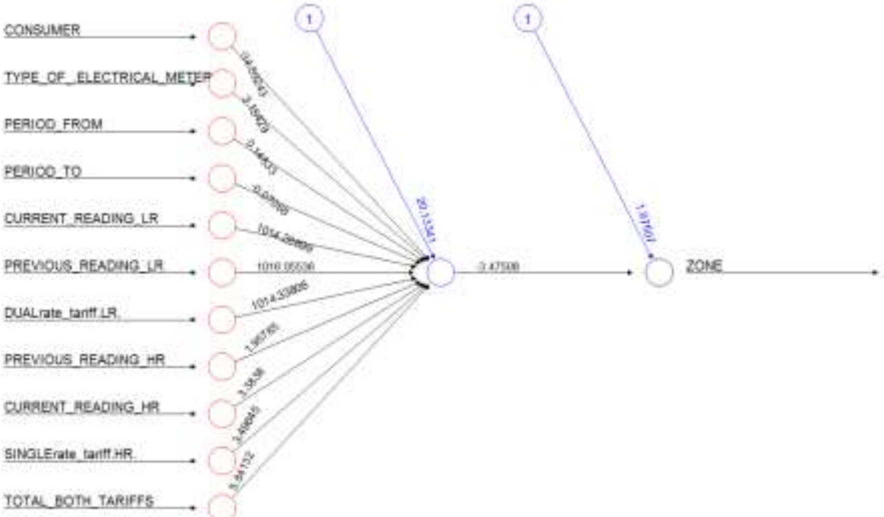

Fig. 3 Neural network diagram created using dependent variable – ‘ZONE’

After creating the network, the results were compared with the data from the test group, and their correlation was reported using the confusion matrix, as shown in Table 5.

Table 5 Confusion Matrix

	prediction	
actual	0	1
0	163965	17154
1	39199	94255

Based on these data, the accuracy of this specific model was calculated using equation 1, as well as its precision (p), sensitivity (r) and F_1-*score*using equations 2, 3 and 4, respectively.

$$\text{Accuracy} = 82\%\,; \qquad p = 90\%; \qquad r = 80\%; \qquad F_1 = 85\%.$$

Given the confusion matrix, other measures can be calculated in addition to the above-mentioned ones. This can also be done using some ready-made software available on the Internet. An example of such software is available at the following address: http://onlineconfusionmatrix.com/ [19]. The measures obtained using this software for the examplesdescribed in sections 3.1 and 3.2 are given in Table 6.

Table 6 Confusion matrix

Measure	Value (classification: *type of electrical meter*)	Value (classification: *zone*)	Derivations
Sensitivity / recall (*r*)	0.7375	0.8071	TPR = TP / (TP + FN)
Specificity	0.7977	0.8460	SPC = TN / (FP + TN)
Precision (*p*)	0.6461	0.9053	PPV = TP / (TP + FP)
Negative Predictive Value	0.8586	0.7063	NPV = TN / (TN + FN)
False Positive Rate	0.2023	0.1540	FPR = FP / (FP + TN)
False Discovery Rate	0.3539	0.0947	FDR = FP / (FP + TP)
False Negative Rate	0.2625	0.1929	FNR = FN / (FN + TP)
Accuracy	0.7776	0.8209	ACC = (TP + TN) / (P + N)
F_1 Score	0.6888	0.8534	F1 = 2TP / (2TP + FP + FN)
Matthews Correlation Coefficient	0.5197	0.6320	TP*TN - FP*FN / sqrt((TP+FP)*(TP+FN)*(TN+FP)*(TN+FN))

4. Conclusion

According to the available literature, data classification via neural networks is one of the most commonly used techniques for processing huge datasets. This technique was used to obtain the results reported in this paper. The results can be further processed for different purposes. Given the fact that a large dataset was dealt with, the results should provide a clear and unambiguous picture of the research.

In the first example, the classification of electricity consumers according to the type of the meter they possess was performed. The accuracy of the predicted data was 77%, and the precision was 65%, which is a satisfactory result.

In the example relating to the predictions according to the zone consumers live in, high precision was achieved as well. The accuracy of 82% and precision of 90% are highly satisfactory. The high precision and favorable F_1-scores indicate that the learning was successfully done, and it can be concluded that the algorithm can provide reliable results regarding the prediction of the zone consumers live in as well.

While processing different examples using the same dataset, with the same dependent and independent variables, some adjustments of the basic network parameters such as the number of hidden layers and threshold, were performed, but it turned out that these parameters did not significantly affect the final values, and therefore the results are not reported herein.

Based on the obtained results, it can be concluded that this learning model can be used reliably enough to predict the type of the electric meter electricity consumers possess, as well as whether they live in an urban or rural area.

This paper presents only a segment of the research and the results obtained, which will provide the basis for further research using some other methods and algorithms, and it will be described in the future papers.

Acknowledgement: *This study was supported by the Serbian Ministry of Education and Science, Project III 44006 and Project III 41007.*

References

[1] U. Ali, C. Buccella and C. Cecati, “Households electricity consumption analysis with data mining techniques”, Department of Information Engineering, Computer Science and Mathematics, University of L’Aquila, Italy, 2016.

[2] D. Shi, J. Guan, J. Zurada and A. Manikas, “A Data-Mining Approach to Identification of Risk Factors in Safety Management Systems”, *Journal of Management Information Systems,* vol. 34, no. 4, pp. 1054–1081, 2017.

[3] M. Blagojević, “Appliance of web mining in education”, Technics and Informatics in Education, Čačak, 2010.

[4] S. Shadroo and M. A. Rahmani, “Systematic survey of big data and data mining in internet of things”, *Computer Networks*, 2018.

[5] G. Prati, L. Pietrantoni and F. Fraboni, “Using data mining techniques to predict the severity of bicycle crashes”, *Accident Analysis & Prevention, Elsevier Ltd,* 2017, pp. 44–54.

[6] M. Carpita, M. Sandri, A. Simonetto and P. Zuccolotto, “Data Mining Applications with R”, Research Center “Data, Methods and Systems” Department of Economics and Management of the University of Brescia, Italy, 2014.

[7] Z. Guo, K. Zhou,X. Zhang, S. Yang and Z. Shao, “Data mining based framework for exploring household electricity consumption patterns: A case study in China context”, *Journal of Cleaner Production, Elsevier Ltd,* 2018.

[8] R. Rathod and R. D. Garg, "Regional electricity consumption analysis for consumers using data mining techniques and consumer meter reading data", *International Journal of Electrical Power & Energy Systems*, vol. 78, pp. 368–374, 2016.

[9] S. K. Barai, "Data mining applications in transportation engineering", Journal Transport, 2003.

[10] C. Da Cunha, B. Agard and A. Kusiak, "Data mining for improvement of product quality", *International Journal of Production Research*, vol. 44, no. 18–19, pp. 4027–4041, 2006.

[11] C. Djeraba, "Data mining from multimedia", *International Journal Parallel Emergent Distributed System*, vol. 22, pp. 405–406, 2007.

[12] S. Xiaogang, "Data Mining Methods and Models", *The American Statistician*, vol. 62, no. 1, pp. 91, 2012.

[13] A. Ghasemi, M. Gitizadeh, "Detection of illegal consumers using pattern classification approach combined with Levenberg-Marquardt method in smart grid", *International journal of electrical power and energy systems*, vol. 99, pp. 363–375, 2018.

[14] S. Ramos; J. M. Duarte; F. J. Duarte; Z. Vale, "A data-mining-based methodology to support MV electricity customers' characterization", Elsevier BV, 2015.

[15] Z. Jiang, R. Lin, F. Yang, "A Hybrid Machine Learning Model for Electricity Consumer Categorization Using Smart Meter Data", *Energies*, vol. 11, p. 2235, 2018.

[16] F. Günther, "neuralnet: Training of Neural Networks", *Stefan Fritsch, The R Journal* , vol. 2/1, 2010.

[17] L. Bing, *Web Data Mining. Exploring Hyperlinks, Contents, and Usage Data*,Secon Edition, Springer, 2011.

[18] G. Ciaburro and B. Venkateswaran, "Neural networks with R", *Packt Publishing Ltd, Birmingham*, 2017.

[19] Software on web address: http://onlineconfusionmatrix.com/, accessed in: October 2018.