DOI: 10.5604/01.3001.0013.9131

PRZEGLĄD STATYSTYCZNY
TOM LXVI – ZESZYT 3 – 2019

Mariusz ŁAPCZYŃSKI[1]
Bartłomiej JEFMAŃSKI[2]

The number of clusters in hybrid predictive models: does it really matter?

Abstract. *For quite a long time, research studies have attempted to combine various analytical tools to build predictive models. It is possible to combine tools of the same type (ensemble models, committees) or tools of different types (hybrid models). Hybrid models are used in such areas as customer relationship management (CRM), web usage mining, medical sciences, petroleum geology and anomaly detection in computer networks. Our hybrid model was created as a sequential combination of a cluster analysis and decision trees. In the first step of the procedure, objects were grouped into clusters using the k-means algorithm. The second step involved building a decision tree model with a new independent variable that indicated which cluster the objects belonged to. The analysis was based on 14 data sets collected from publicly accessible repositories. The performance of the models was assessed with the use of measures derived from the confusion matrix, including the accuracy, precision, recall, F-measure, and the lift in the first and second decile. We tried to find a relationship between the number of clusters and the quality of hybrid predictive models. According to our knowledge, similar studies have not been conducted yet. Our research demonstrates that in some cases building hybrid models can improve the performance of predictive models. It turned out that the models with the highest performance measures require building a relatively large number of clusters (from 9 to 15).*

Keywords: hybrid predictive model, k-means algorithm, decision trees

JEL Classification: C10, C18, C52

1. INTRODUCTION

The aim of this paper is to check to what extent the number of clusters affects the quality of predictive models which combine decision trees with cluster analysis (centre-based algorithm). The concept of the hybridization of the two methods is not new – it was already applied to customer relationship management (Chu et al., 2007; Bose and Chen, 2009; Li et al., 2011), the analysis of the Internet users' patterns of behaviour (Łapczyński and Surma, 2012), medical sci-

[1] Cracow University of Economics, College of Management and Quality Sciences, Department of Market Analysis and Marketing Research, ul. Rakowicka 27, 31-510 Cracow, Poland, corresponding author – e-mail: lapczynm@uek.krakow.pl, ORCID: https://orcid.org/0000-0002-4508-7264.

[2] Wroclaw University of Economics and Business, Faculty of Economics and Finance, Department of Econometrics and Informatics, ul. Nowowiejska 3, 58-500 Jelenia Góra, Poland, ORCID: https://orcid.org/0000-0002-0335-0036.

ences (Khan and Mohamudally, 2011; Shouman et al., 2012), petroleum geology (Ferraretti et al., 2011) and the detection of computer anomalies (Gaddam et al., 2007). Hybrid models differ from ensemble models in that they combine two different analytical tools. In the case of ensemble models, the commonly used procedures include the random forest, rotation forest, and boosted trees. Hybrid models, on the other hand, are referred to as cascade models, cross-algorithm ensembles, and two-stage classification (Łapczyński and Jefmański, 2013).

We decided to combine a popular decision tree algorithm CART (classification and regression trees) with the *k*-means algorithm. The hybridization process was tested on 14 data sets downloaded from publicly accessible online repositories. Each of them had a qualitative dependent variable with two or more categories, and a set of independent variables presented on various measurement scales. In addition, we calculated 4 cluster validity measures. However, our primary objective was to analyse hybrid models based on 2 to 20 clusters.

The paper consists of 4 sections. Section 2 encompasses a brief description of the employed analytical tools and the method for building a hybrid model. It also presents the characteristics of data sets and the description of the process of data preparation. Section 3 discusses the results of the study and provides the assessment of the quality of hybrid models based on five performance measures. The authors also explain there why hybrid models demonstrate a higher predictive power for some of the data sets than for the others. The conclusions and recommendations are provided in the last section.

2. THE CHARACTERISTICS OF A HYBRID MODEL AND EMPLOYED DATA SETS

2.1 CART – *k*-MEANS HYBRID MODEL

A decision tree is a commonly used analytical tool for data mining. The analysis utilises the CART algorithm, developed by Breiman et al. (1984). This tool demonstrates great flexibility in terms of the measurement scale of independent variables. It does not have such a great predictive power as ensemble models, but it enables creating a set of rules according to an 'if ... then ...' formula, which is easy to understand for managers with no mathematical background. The analysis adopts the CART algorithm where equal a priori probabilities and equal misclassification costs have been assumed. A minimum number of cases in tree leaves is placed at the level of 2% of the training set.

A cluster analysis with the use of the *k*-means algorithm is a commonly adopted approach in statistical exploratory analyses as well as in data mining. The algorithms applied most frequently to such type of research include the Lloyd, the MacQueen and the Hartigan and Wong algorithms (Everitt et al., 2011). These algorithms are relatively easy to use, have a large calculating potential and require relatively little computer memory compared to other clustering

algorithms. Research studies do not identify the best cluster analysis algorithm. The choice of a specific algorithm depends on the structure of a data set, its size, the number of analysed variables, etc. Due to large sizes of our data sets, we employed the Lloyd algorithm (implemented in the Statistica software). It is one of the most commonly used data mining algorithms. Its popularity stems from three main reasons (Lloyd, 1982):

- Minimizing an objective function is relatively easy and intuitive,
- The algorithm is simple, effective and often leads to optimal solutions,
- The results of the analysis are easily interpretable.

The characteristic feature of the methods for optimizing the initial partition of objects is an a priori determination of the number of clusters. One of the ways to conduct an analysis in this area is to estimate this number by means of the classification quality measures. However, as emphasized by Everitt et al. (2011), the selection of the optimal number of clusters should be done on the basis of the synthesis of the results obtained with the help of other methods. Such a procedure is recommended e.g. due to the fact that each method is based on predefined assumptions referring to the structure of classes, which are not always satisfied. Therefore, in our analysis, we applied several measures that are frequently implemented in empirical research studies and are available in the R package clusterSim: the Calinski-Harabasz index, the Krzanowski-Lai index, the Davies and Bouldin index, the Gap Statistic (Walesiak and Dudek, 2011). The hybridization procedure consists of the following steps:

1. The indication of the qualitative dependent variable and the set of independent variables within the data set,
2. The selection of quantitative independent variables from the set of independent variables and their application to building clusters, and subsequently the replacement of all the quantitative variables in the predictive model by the new variable informing about cluster membership,
   a. Subjective determination of the number of clusters or determination of the number by means of any cluster-validity measure,
   b. The reduction of the number of quantitative independent variables using the Random Forest if the number of such variables exceeds 15; more specifically, the selection of 15 variables on the basis of the variable-importance ranking,
3. The construction of a decision tree model by means of all qualitative independent variables and the new qualitative independent variable created in step 2.

We sequentially combined both analytical tools, thus creating a hybrid CART – k-means model. In the first step of the procedure, we created clusters on the basis of quantitative independent variables from the data set. The number of clusters could not exceed 15 (Blattberg et al., 2008). If a data set consisted of a larger number of quantitative variables, it was necessary to select 15 of them. This selection was carried out with the help of the Random Forest, which is

a method that allows creating a variable-importance ranking. The 15 variables thus selected were those most strongly related to the dependent variable. In the second step, a decision tree model was built, which comprised of qualitative independent variables and the new variable providing information about the cluster membership. The original quantitative variables were not used in the analysis.

The cluster analysis determined 2–20 clusters, which implied that the analysis of each data set yielded 19 different hybrid models. Setting the maximum number of clusters to 20 was our subjective choice. This value was higher than the maximum number of clusters indicated by the cluster validity measures used in the study. All quantitative variables used in the cluster analysis were standardized using the z-score formula ((value-mean)/standard deviation). We also calculated cluster validity measures, but their values did not determine the optimal number of clusters.

Additionally, a decision tree model based on the entire non-transformed set of independent variables (both categorical and numerical) was built for each data set (the so-called base tree). The decision tree is characterised by the following parameters: split rule – Gini measure, equal misclassifications costs, equal a priori probabilities, minimal number of cases in a parent node (5% of training set), minimal number of cases in a leaf (2% of training set) and maximum depth of the tree (15 levels). Its performance was a reference point for comparable hybrid models. The number of predictive models used for the purposes of the analysis totalled 280.

2.2 DESCRIPTION OF DATA SETS

Most of the data sets used in the experiment come from a well-known UCI machine learning repository (Asuncion and Newman, 2007). Table 1 provides information on the name of the data set, the number and type of independent variables, the number of categories of the dependent variable and the number of cases. Originally, this repository was intended to select data sets relating to the analytical CRM, database marketing and other business analytical areas. It was also important that the dependent variable was binary. Unfortunately, during the collection of data, it turned out that this type of data is confidential and is very rarely available in publicly-accessible online repositories. Ultimately, we decided to choose data sets with a varying number of cases (from 208 to 50,000), different numbers of dependent variable categories (from 2 to 10) and different numbers and types of independent variables (from 4 to 111). According to our intentions, this diversity was to ensure more reliable testing of hybrid models.

Each data set was divided into a training set (70 %), and a test set (30 %). The variables for which the missing data exceeded 10 percent, and the instances for which the missing data exceeded 50 percent, were excluded from the analysis. In the remaining cases, the missing data were substituted for by mean or modal values. We decided to replace the missing data by the simplest meth-

ods to eliminate their possible impact on the quality of the predictive models. The variables possessing unique values (ID, phone number, dates) were not analysed.

The models were assessed on the basis of measures calculated with the use of the misclassification matrix: accuracy ((TP + TN) / (TP + FP +TN + FN)), recall (TP / (TP + FN)), precision (TP / (TP + FP)) and F-measure ((2 × precision × recall) / (precision + recall)). The acronyms used in the formulas come from the confusion matrix and represent true positive (TP), true negative (TN), false positive (FP) and false negative (FN). Additionally, the lift measure in the first and second decile of test set was calculated. The lift is the ratio of the response rate in a decile to the average response rate (in the whole data set).

TABLE 1. THE CHARACTERISTICS OF DATA SETS

Data set	Number of independent variables	Number of categories of dependent variable	Number of cases
(D1) Bank Marketing	6 numerical and 8 categorical	2	45211
(D2) German Credit	7 numerical and 13 categorical	2	1000
(D3) Insurance Company	23 numerical and 62 categorical	2	5822
(D4) Churn	15 numerical and 5 categorical	2	5000
(D5) KDD 2009 (preprocessed)	3 numerical and 18 categorical	2	50000
(D6) CINA Marketing	3 numerical and 108 categorical	2	16033
(D7) Australian Credit	6 numerical and 8 categorical	2	690
(D8) Banknote	4 numerical	2	1372
(D9) Heart (Statlog)	5 numerical and 8 categorical	2	270
(D10) Ionosphere	34 numerical	2	352
(D11) Pendigits	16 numerical	10	10992
(D12) Image Segment	14 numerical and 4 categorical	7	2310
(D13) Sonar	60 numerical	2	208
(D14) Vehicle	18 numerical	4	846

S o u r c e: own compilation.

3. RESULTS OF EXPERIMENT

Table 2 presents the selected performance measures for all data sets. The measure for the best hybrid model is placed in front of the bracket, whereas the measure for the base tree inside the bracket. In some cases, the difference was

in the third decimal place, but it is not visible after rounding the results. Data presented in the table indicate that in 8 out of 14 data sets (D1, D2, D6-D9, D11, and D13) a hybrid approach was more effective than unmodified decision tree, considering all the measures.

TABLE 2. INCREASED VALUES OF PERFORMANCE MEASURES IN HYBRID MODELS AS COMPARED WITH THE BASE TREE

Data set	Accuracy	Precision	Recall	*F*-measure	Lift 10%	Lift 20%
D1	0.87 (0.77)	0.45 (0.29)	0.78 (0.71)	0.46 (0.42)	4.23 (3.59)	3.20 (2.55)
D2	0.67 (0.66)	0.50 (0.49)	0.84 (0.77)	0.61 (0.60)	1.95 (1.50)	1.76 (1.50)
D3	0.71 (0.59)	0.14 (0.10)	0.75 (0.71)	0.22 (0.18)	3.17 (3.17)	2.51 (2.51)
D4	0.82 (0.86)	0.40 (0.49)	0.77 (0.78)	0.50 (0.60)	3.94 (4.21)	2.94 (3.59)
D5	0.69 (0.65)	0.10 (0.10)	0.73 (0.44)	0.16 (0.16)	1.85 (1.85)	1.45 (1.39)
D6	0.92 (0.90)	0.82 (0.76)	0.90 (0.87)	0.84 (0.81)	3.62 (3.11)	3.62 (3.11)
D7	0.89 (0.87)	0.90 (0.86)	0.91 (0.85)	0.88 (0.85)	2.17 (2.07)	2.09 (2.07)
D8	0.98 (0.91)	0.96 (0.90)	1.00 (0.92)	0.98 (0.91)	2.08 (2.06)	2.08 (2.06)
D9	0.82 (0.70)	0.77 (0.60)	0.80 (0.79)	0.77 (0.68)	2.23 (2.03)	2.10 (2.03)
D10	0.88 (0.88)	1.00 (1.00)	0.78 (0.64)	0.80 (0.78)	2.92 (2.92)	2.92 (2.92)
D11	0.85 (0.80)	0.97 (0.94)	0.98 (0.81)	0.96 (0.87)	8.84 (7.25)	4.83 (3.47)
D12	0.92 (0.92)	1.00 (0.96)	1.00 (1.00)	0.99 (0.98)	7.07 (6.79)	4.97 (1.00)
D13	0.85 (0.74)	0.89 (0.71)	0.93 (0.80)	0.85 (0.75)	2.07 (1.03)	2.07 (1.45)
D14	0.63 (0.46)	0.85 (0.85)	0.98 (0.84)	0.85 (0.85)	3.63 (3.57)	3.50 (3.50)

S o u r c e: own calculations.

Table 3 presents the minimal number of clusters which we needed to build the best hybrid model in our computer experiment. We intended to create the smallest possible number of clusters, which would facilitate their descriptions. Unfortunately, approximately 60% of the models yielded 10 or more clusters. Moreover, it turned out that the optimal number of clusters indicated by cluster validity measures did not provide best solutions. Also, when the number of clusters reached 20, it became possible that a higher value of performance measures could have been obtained for a larger number of clusters.

TABLE 3. MINIMUM NUMBER OF CLUSTERS IN THE BEST HYBRID MODELS

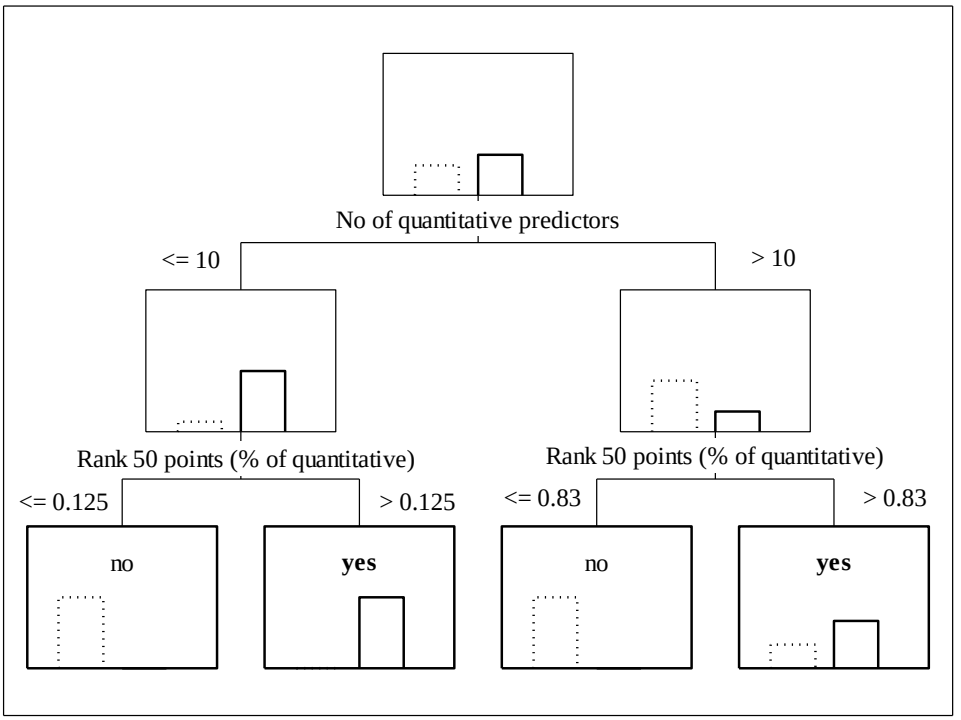
Data set	Accuracy	Precision	Recall	*F*-measure	Lift 10%	Lift 20%
D1	9	9	14	3	4	3
D2	3	12	10	16	17	9
D3	10	10	17	10	2	2
D4	7	7	14	17	8	14
D5	13	2	10	5	2	18
D6	20	20	2	20	20	20
D7	18	16	14	18	2	15
D8	20	20	13	20	11	11
D9	15	15	12	12	16	6
D10	13	18	9	13	10	18
D11	19	16	2	19	16	20
D12	18	3	4	4	2	3
D13	6	12	4	6	5	5
D14	19	16	2	18	19	18

S o u r c e: own compilation.

Subsequently, we investigated the reasons for the successes and failures of hybrid models. For this purpose, we employed the variable-importance ranking of the CART algorithm, which can assign from 0 to 100 points to all independent variables (Breiman et al., 1984). The higher the ranking position, the stronger the relationship between the predictor and the dependent variable.

In the next step we checked the relationship between the number of quantitative predictors with the largest number of assigned points and the quality of hybrid models. It was assumed that a strong relation between quantitative variables and the dependent variable indicates a strong relation between clusters and the dependent variable. Some other hypothetical success factors included the number of qualitative independent variables, the number of quantitative independent variables, the number of cases in the data set, the number of categories of the dependent variable and the difference in the numbers of observations among the categories of dependent variables (the latter variable provides information on the imbalance class problem). The next step involved building a decision tree in which the binary dependent variable assumed two values: 1 for the success of a hybrid model, and 0 for its failure. The set of independent variables comprised of all the above-mentioned determinants of the quality of hybrid models.

Figure 1. Decision tree classifying the result of the hybrid model
(binary dependent variable 'success' with two categories: yes / no)



S o u r c e: own compilation.

Figure 1 presents a CART decision tree model with 4 terminal nodes which encompass the best hybrid models (the leaves bear a 'yes' label). The highest quality of hybrid models was recorded for data sets where:

- the number of numerical independent variables was equal to or smaller than 10, and more than 12.5% of numerical predictors were assigned over 50 points (6 models),
- the number of numerical independent variables was larger than 10, and more than 83% of numerical predictors were assigned over 50 points (2 models).

In simplified terms, it can be stated that if a data set comprises 10 or fewer quantitative independent variables, hybrid models are more effective than a base tree. Such a result is obtained for 6 out of 7 sets. The success of a hybrid approach may result from the manner of dividing the classification tree. When quantitative predictors are used, the number of possible splits of nodes is equal to or smaller than n, where n indicates the number of predictor values. In the case of qualitative predictors, the number of possible splits is much larger, amounting to $2^{n-1}-1$, where n indicates the number of predictor categories. A larger number of possible splits can lead to a greater number of possible, and sometimes better, solutions.

TABLE 4. ERROR RATES AND ERROR RATES AFTER 10-FOLD CV (IN BRACKETS) FOR THE BEST HYBRID MODELS ESTIMATED ON THE BASIS OF THE TRAINING SET

Data set	D1	D2	D6	D7	D8	D9	D11	D13
base tree	0.2272 (0.2273)	0.2400 (0.3756)	0.0964 (0.0977)	0.1304 (0.1808)	0.0698 (0.0908)	0.1005 (0.1771)	0.1972 (0.2172)	0.0616 (0.2803)
2 clusters	0.2817 (0.2733)	0.2871 (0.3658)	0.1209 (0.1213)	0.1180 (0.1542)	0.4303 (0.4359)	0.0899 (0.1486)	0.7969 (0.8015)	0.3630 (0.3561)
3 clusters	0.1875 (0.1889)	0.2743 (0.3669)	0.1209 (0.1213)	0.1180 (0.1542)	0.3292 (0.3575)	0.1005 (0.1829)	0.7011 (0.7026)	0.3014 (0.2955)
4 clusters	0.2050 (0.1846)	0.2943 (0.3903)	0.0997 (0.1003)	0.1139 (0.1610)	0.2781 (0.2736)	0.1058 (0.1676)	0.6106 (0.6120)	0.2329 (0.2348)
5 clusters	0.1425 (0.1570)	0.2971 (0.3793)	0.0943 (0.0948)	0.1180 (0.1497)	0.1719 (0.1725)	0.1005 (0.1977)	0.5537 (0.5539)	0.3014 (0.3030)
6 clusters	0.1411 (0.1375)	0.2929 (0.3756)	0.1209 (0.1164)	0.1097 (0.1545)	0.1937 (0.1986)	0.1058 (0.1839)	0.4574 (0.4594)	0.2055 (0.2576)
7 clusters	0.1299 (0.1294)	0.3000 (0.3683)	0.0831 (0.0871)	0.1284 (0.1549)	0.1833 (0.1884)	0.0847 (0.1552)	0.4029 (0.4053)	0.2123 (0.2955)
8 clusters	0.1406 (0.1597)	0.2714 (0.3756)	0.1011 (0.1022)	0.1180 (0.1629)	0.1552 (0.1703)	0.0952 (0.1609)	0.3253 (0.3283)	0.2192 (0.2424)
9 clusters	0.1249 (0.1287)	0.2871 (0.3699)	0.0981 (0.0991)	0.1201 (0.1606)	0.1687 (0.1646)	0.1058 (0.1607)	0.2629 (0.2660)	0.1644 (0.1667)
10 clusters	0.2324 (0.2441)	0.3400 (0.3962)	0.0835 (0.0839)	0.1139 (0.1640)	0.1208 (0.1215)	0.0794 (0.2326)	0.2302 (0.2325)	0.1781 (0.2045)
11 clusters	0.1584 (0.2514)	0.2743 (0.3558)	0.0950 (0.0964)	0.1200 (0.1558)	0.1156 (0.1283)	0.0952 (0.2081)	0.2798 (0.2814)	0.1781 (0.1818)
12 clusters	0.2263 (0.2367)	0.2843 (0.3443)	0.0793 (0.0864)	0.1284 (0.1587)	0.1053 (0.1056)	0.0741 (0.2289)	0.2031 (0.2052)	0.1849 (0.2045)
13 clusters	0.2056 (0.1965)	0.2643 (0.3742)	0.0793 (0.0836)	0.1180 (0.1490)	0.0531 (0.0579)	0.0794 (0.1786)	0.2085 (0.2104)	0.1781 (0.2348)
14 clusters	0.2605 (0.2413)	0.3257 (0.3903)	0.0850 (0.0856)	0.1325 (0.1603)	0.1146 (0.1169)	0.0847 (0.1697)	0.2972 (0.3015)	0.1986 (0.2424)
15 clusters	0.1724 (0.2004)	0.2857 (0.3956)	0.0809 (0.0815)	0.1014 (0.1621)	0.0260 (0.0284)	0.0741 (0.1905)	0.2250 (0.2240)	0.1849 (0.2061)
16 clusters	0.1642 (0.1697)	0.2971 (0.3664)	0.0812 (0.0858)	0.1284 (0.1746)	0.0448 (0.0477)	0.0952 (0.2651)	0.1860 (0.1869)	0.1644 (0.1985)
17 clusters	0.1965 (0.2028)	0.2686 (0.3836)	0.0830 (0.0835)	0.1180 (0.1575)	0.0323 (0.0318)	0.0847 (0.1863)	0.1469 (0.1471)	0.1712 (0.2308)
18 clusters	0.1787 (0.1755)	0.2442 (0.3506)	0.0796 (0.0862)	0.1221 (0.1713)	0.0615 (0.0670)	0.0847 (0.2048)	0.1457 (0.1449)	0.1712 (0.1756)
19 clusters	0.2134 (0.2285)	0.2786 (0.3831)	0.0790 (0.0796)	0.1118 (0.1475)	0.0437 (0.0465)	0.0899 (0.2073)	0.1522 (0.1522)	0.1918 (0.2424)
20 clusters	0.2005 (0.2080)	0.2571 (0.3642)	0.0781 (0.0785)	0.1201 (0.1558)	0.0177 (0.0193)	0.1164 (0.2201)	0.1448 (0.1465)	0.1438 (0.1742)

S o u r c e: own compilation.

Table 4 presents error rates and error rates after 10-fold cross validation (in brackets). The figures refer only to those data sets for which the hybrid models provided the best performance measures. Both error rates were estimated using the training set, because only that set contained variables informing about the class membership. The test set was used twice during the model evaluation. Firstly, the cluster's membership was predicted on the basis of quantitative variables. Subsequently we predicted the class of variable Y. The comparison of both values made it possible to assess the stability of the results, although in this case it was limited to the training set.

4. CONCLUSIONS

Building hybrid models which combine decision tree algorithms with cluster analysis can, in some cases, improve the performance of predictive models. Prior to starting analytical research, it may be worthwhile checking the relationships between independent variables and the dependent variable. This refers in particular to the number of quantitative predictors and their position in the variable-importance ranking. The process of building clusters cannot rely on cluster validity measures, because they indicate different numbers of clusters, and do not always guarantee good quality of hybrid models.

The weakness of this approach is reflected by a large number of clusters in the best hybrids. The average number of clusters in the hybrid predictive model providing the highest accuracy value was 14. For the remaining best-performance measures, the average number of clusters was: recall (9 clusters), precision (15 clusters), F-measure (14 clusters), and lift in both deciles (11 clusters). This had a negative impact on the possible interpretation of a model, making a hybrid approach similar to a black box, which we intended to avoid. Our intention was to build a model that would have higher predictive power and at the same time would not lose the properties of decision trees, i.e. would yield a set of easily interpretable "if ... then ..." rules.

Undoubtedly, the limitation of this analytical experiment was the lack of cross-validated error rates that would be estimated on the basis of the entire data set. This made it impossible to assess the stability of the results. Moreover, we are aware that our approach should have been compared with univariate optimal binning methods. This is a popular method for transforming quantitative variables into qualitative ones.

It should be noted that despite the double use of a test set (firstly, when objects were assigned to clusters, and again in the process of deployment the decision tree model), performance measures assumed higher values than in the base tree. Furthermore, a higher quality of hybrid models was achieved, despite the sensitivity of cluster analysis to outliers or the risk resulting from finding artefactual solution (lack of natural clusters in data). These promising results encourage further research in this area. They could be extended by

utilising a larger number of data sets or the employment of different decision tree algorithms (C4.5 or CHAID) or cluster analysis algorithms (Mac Queen's or Hartigan and Wong's).

REFERENCES

Asuncion A., Newman D., (2007), UCI machine learning repository, http://archive.ics.uci.edu.

Blattberg R., Kim B. D., Neslin S., (2008), *Database Marketing – Analyzing and Managing Customers, 1st ed.*, Springer, New York. DOI: 10.1007/978-0-387-72579-6.

Bose I., Chen X., (2009), Hybrid Models Using Unsupervised Clustering for Prediction of Customer Churn, *Journal of Organizational Computing and Electronic Commerce*, 19(2), 133–151, DOI: 10.1080/10919390902821291.

Breiman L., Friedman J., Olshen R., Stone C., (1984), *Classification and Regression Trees, 1st ed. Wadsworth statistics / probability series*, Wadsworth Publishing Company, Belmont, California.

Chu B. H., Tsai M. S., Ho C. S., (2007), Toward a Hybrid Data Mining Model for Customer Retention, *Knowledge-Based Systems*, 20(8), 703–718. DOI: 10.1016/j.knosys.2006.10.003.

Everitt B., Landau S., Leese M. D. S., (2011), *Cluster Analysis, 5th ed. Wiley Series in Probability and Statistics*, John Wiley & Sons, Chichester, West Sussex. DOI: 10.1002/9780470977811.

Ferraretti D., Lamma E., Gamberoni G., Febo M., Di Cuia R., (2011), Integrating Clustering and Classification Techniques: A Case Study for Reservoir Facies Prediction, [in:] Ryżko D., Gawrysik P., Rybiński H., Kryszkieiwcz M., *Emerging Intelligent Technologies in Industry*, Springer, Berlin, Heidelberg, 21–34. DOI: 10.1007/978-3-642-22732-5_3.

Gaddam S., Phoha V., Balagani K., (2007), K-means + ID3: A Novel Method for Supervised Anomaly Detection by Cascading K-means Clustering and ID3 Decision Tree Learning Methods, *IEEE Transactions on Knowledge and Data Engineering*, 19(3), 345–354.DOI: 10.1109/TKDE.2007.44.

Khan D., Mohamudally N., (2011), An Integration of *k*-means and Decision Tree (ID3) Towards a More Efficient Data Mining Algorithm, *Journal of Computing*, 3(12), 76–82, https://sites.google.com/site/journalofcomputing/volume-3-issue-12-december-2011.

Łapczyński M., Jefmański B., (2013), Impact of Cluster Validity Measures on Performance of Hybrid Models Based on K-means and Decision Trees, [in:] Perner P., (ed.), *Advances in Data Mining*, Ibai Publishing, Fockendorf, 153–162.

Łapczyński M., Surma J., (2012), Hybrid Predictive Models for Optimizing Marketing Banner Ad Campaign in Online Social Network, [in:] Stahlbock R., (ed), *Proceedings of the 2012 International Conference on Data Mining (DMIN)*, CSREA Press, Las Vegas, Nevada, 140–146.

Li Y., Deng Z., Qian Q., Xu R., (2011), Churn Forecast Based on Two-step Classification in Security Industry, *Intelligent Information Management*, 3(4), 160–165. DOI: 10.4236/iim.2011.34019.

Lloyd S., (1982), Least Squares Quantization in PCM. *IEEE Transactions on Information Theory*, 28(2), 129–137, Institute of Electrical and Electronics Engineers (IEEE). DOI: 10.1109/TIT.1982.1056489.

Shouman M., Turner T., Stocker R., (2012), Integrating Decision Tree and K-Means Clustering with Different Initial Centroid Selection Methods in the Diagnosis of Heart Disease Patients, [in:] Stahlbock R., (ed), *Proceedings of the 2012 International Conference on Data Mining (DMIN)*, CSREA Press, Las Vegas, Nevada, 24–30.

Walesiak M., Dudek A., (2011), clusterSim: Searching for Optimal Clustering Procedure for a Data Set, https://cran.r-project.org/web/packages/clusterSim. R package version 0.47–3.