STUDIES IN LOGIC, GRAMMAR
AND RHETORIC 37 (50) 2014
DOI: 10.2478/slgr-2014-0022

# Number of Clusters and the Quality of Hybrid Predictive Models in Analytical CRM

**Mariusz Łapczyński**[1], **Bartłomiej Jefmański**[2]

[1] Department of Market Analysis and Marketing Research, Cracow University of Economics, Poland, lapczynm@uek.krakow.pl

[2] Department of Econometrics and Computer Science, Wroclaw University of Economics, Poland, bartlomiej.jefmanski@ue.wroc.pl

**Abstract.** Making more accurate marketing decisions by managers requires building effective predictive models. Typically, these models specify the probability of customer belonging to a particular category, group or segment. The analytical CRM categories refer to customers interested in starting cooperation with the company (acquisition models), customers who purchase additional products (cross- and up-sell models) or customers intending to resign from the cooperation (churn models). During building predictive models researchers use analytical tools from various disciplines with an emphasis on their best performance. This article attempts to build a hybrid predictive model combining decision trees (C&RT algorithm) and cluster analysis (k-means). During experiments five different cluster validity indices and eight datasets were used. The performance of models was evaluated by using popular measures such as: accuracy, precision, recall, G-mean, F-measure and lift in the first and in the second decile. The authors tried to find a connection between the number of clusters and models' quality.

*Keywords*: hybrid predictive models, analytical CRM, decision trees, k-means

## 1. Introduction

The process of making marketing decisions in the company refers to consumer decision making and managerial decision making, among which one can distinguish several areas including: advertising, sales promotions, sales management, competition, customer relationship marketing (CRM) (Wierenga, 2008). In order to improve the decision making process managers utilize predictive models dating back to the 60s of the last century, when the microeconomic approach was first used to solve marketing problems. Contemporary marketing models that are related to analytical CRM are based on the popular ACURA concept (acquire, cross-sell, up-sell, retain, advocacy) and are closely related to customers' lifecycles (Christopher, Payne, Ballantyne, 2002). Predictive models are built by using a wide range of analytical tools which have their roots in microeconomics, mathematics, statistics, econometrics, and data mining.

The construction of predictive models in customer relationship management refers to each stage in the customer's lifecycle, i.e. the customer acquisition, development and retention. In these areas one frequently applies supervised methods such as decision trees, neural networks, Random Forest, boosted trees, logistic regression, discriminant analysis, etc. Generally, the analyst's target is the construction of such a model that will in the best possible way anticipate the customer's sense of belonging to a particular category of the dependent variable (potential customer, potential churner, etc.). Constructing models is associated with the analytical CRM, whereas conducting marketing activities on their basis is comprised in the operational CRM. Sometimes the terms back-office systems and front-office systems are applied here.

This article aims at verifying in what manner the measures indicating the optimal number of clusters influence the quality of hybrid predictive models combining the k-means algorithm with classification and regression trees (C&RT)[1]. Combining the clustering analysis with decision trees has recently become a popular method of increasing the performance of predictive models. Research studies covering this area pertain to numerous disciplines, such as customer relationship management, web usage mining, medical sciences, petroleum geology, anomalies in computer networks, etc. The inspiration to undertake the subject came from the successful experiment referring to (Łapczyński, Surma, 2012, p. 140–146) profiling users clicking on the banner ad of a cosmetics company, which was placed on a social networking website which was popular in Poland.

One may notice that the construction of predictive models is more and more frequently accompanied by an attempt to combine analytical tools of the same type and create the so-called ensemble models, also referred to as committees. There are also attempts combining various methods, which are described with the terms 'hybrid', 'two stage classification', 'cascade classification' or 'cross-algorithm ensemble'. In numerous cases such combined attempts permitted to achieve a better performance.

The authors of this article have decided to conduct an experiment consisting in combining the k-means algorithm with decision trees (C&RT). While creating clusters they implemented 5 different cluster validity measures (the Calinski-Harabasz index, the Krzanowski-Lai index, the Davies-Bouldin index, the Hartigan index, and the gap statistic) and observed in what manner the number of clusters influences the performance of the model. The analysis was carried out on 8 data sets collected from publicly accessible repositories. The dependent variable in each dataset possessed two categories, and the set itself as much as possible pertained to the broadly understood marketing activities of a company.

The second section provides a brief review of the literature in which clustering was combined with decision trees during the construction of predictive models. The third section contains a description of model hybridization, characteristics of cluster validity indices as well as characteristics of the implemented datasets. Section IV will present the results of the experiment alongside with the performance evaluation. Section V contains the summary and proposals regarding the successive experiments in this area.

## 2. Examples of hybrid predictive models based on clustering and decision trees

Combining clustering with decision trees for building predictive models has long been of interest to many researchers. It seems that in the field of marketing churn modelling has become popular in recent years. Some authors (Bose, Chen, 2009, p. 133–151) combined the results obtained from clustering algorithms (k-means, k-medoid, self-organizing maps (SOM), fuzzy c-means and Balanced Iterative Reducing and Clustering using Hierarchies (BIRCH)) with the results obtained from the decision tree (C5.0) with boosting. Their goal was to predict the customer churn. During the clustering phase Dunn's index was utilized for the identification of the optimal number of clusters. Two methods of hybridization were examined. In the first approach a new variable was added whose categories informed about the cluster membership while building the decision tree. In the second approach different decision trees were built separately for each cluster. Subsequently the top decile lift was used as a performance measure of hybrid models. It turned out that SOM combined with C5.0 provided the best results in terms of short-term prediction while BIRCH joined with C5.0 provided the best results in terms of long-term prediction. In general, using an additional variable while building decision trees outperformed building separate models in each cluster.

Chu et al (Chu, Tsai, Ho, 2007, p. 703–718) proposed a hybrid model to predict churning in the area of customer relationship management. C5.0 decision trees and Growing Hierarchical Self-Organizing Map (GHSOM) were combined. In the first step the predictive model was constructed on the basis of such independent variables as: defection history, de-activation data, payment history, usage patterns etc. In the second step GHSOM was applied to build four disjoint clusters containing churners. Clustering was based on 9 variables that were recognized by the decision tree to appear in the churn model. In the final step different retention policies were recommended for each cluster, however, no comparison of performance measures was delivered.

Another slightly differing approach of constructing hybrid models was called 'the model of two-step classification' and was proposed (Li, Deng, Qian, Xu, 2011, p. 160–165) as an alternative approach for churn modelling in the security industry. In the first stage of the procedure self-organizing maps (SOM) were used to divide customers into 9 clusters. Clustering was based on such indices as stocks speculation years, capital scale, customer holding ratio, customer transaction times, average commission, etc. In the next step the authors chose the largest segment with the highest churn rate (12.38% versus 5.2% in the entire dataset). Moreover, the number of disloyal clients in that particular cluster accounted for almost 65% of the total churn in the entire dataset. In the last stage a decision tree model was built that provided a high accuracy of classification, which reached nearly 95%.

The combination of the K-means algorithm and decision trees (ID3) was also used in the classification of anomalies in computer networks (Gaddam, Phoha, Balagani, 2007, p. 345–354). In the first step the authors grouped behaviors in a computer network into disjoint clusters in such a way that each cluster consisted of normal and anomaly instances. In the second step there was a decision tree model built separately for each cluster. The approach of joining these two machine learning algorithms was called 'the cascade' one. During the testing phase, each observation was assigned to the cluster where the distance to its centroid was the lowest. Subsequently the instance was classified as 'anomaly' or 'normal' by using two methods. It belonged to the class 'anomaly' when the probability of that class in the cluster was greater than 0.5 (the Bayes rule) or greater than the threshold set by analysts (the Threshold rule). At the same time the instance was classified by using "if...then..." rules from the decision tree. As a consequence, the anomaly score matrix was obtained. In the last step of the procedure the classification was based on Nearest-Consensus Rule or Nearest-Neighbor Rule. In the first case the object (behavior) was assigned to the class for which a consensus between two algorithms had been reached, even if it was not the best cluster (with the lowest distance to the centroid). In the other case the object belonged to the class indicated by the decision tree built in the nearest candidate cluster. To compare the performance of models six measures were used: true positive rate (also referred to as recall), false positive rate, precision, accuracy, F-measure, and area under ROC curve. The cascade model was compared with K-means and ID3 separately. The results of experiments on three data sets are not explicit, however, it seems that a combined approach delivers higher values of precision and F-measure.

Hybrid models combining the cluster analysis with decision trees are also referred to as ‘integrated’ ones. An example of such an approach was an attempt at predicting heart diseases (Shouman, Turner, Stocker, 2012, p. 24–30), in which the dataset was divided into clusters by using k-means algorithm, and afterwards one decision tree model was built for each cluster. The authors investigated the impact of different initial centroid selection methods on the performance of decision trees. The number of clusters ranged from 2 to 5 and the applied methods were as follows: inlier method, outlier method, range method, random attribute method, and random row method.

The evaluation of predictive models performance was based on sensitivity, specificity, and accuracy. The results of the experiment indicate the superiority of the inlier method in the case of the smallest number of clusters equal to 2. The small number of clusters was explained by a small number of observations. Moreover, the integrated predictive model achieved a higher accuracy (83.9%) as compared to a simple decision tree (78.91%) or a bagging algorithm (81.41%).

The term ‘integration’ in the context of combining clustering with classification was also used by Kumar and Rathee (2011, p. 29–33) and Ferraretti et al (Ferraretti, Lamma, Gamberoni, Febo, Di Cuia, 2011, p. 21–34). In the first study, k-means algorithm was joined with J4.8 decision tree algorithm in such a way that 3 different tree models were built in 3 different clusters. The experiment was based on the popular ‘iris’ dataset. The hybrid approach outperformed a simple decision tree as far as overall accuracy is concerned (98.77% versus 95.33% in the case of a simple tree). Additionally, in each cluster higher values of sensitivity, specificity, precision and F-measure were obtained when comparing these results with a simple decision tree model. The other study refers to petroleum geology (characterization of reservoirs) and combines hierarchical clustering with several classification algorithms including J4.8. In the first phase of the procedure geologists identified 8 clusters on the basis of a dendrogram in such a way that they referred to different types of rocks. In the second phase this new independent variable informing about class membership was added to the dataset and several classification techniques were applied. In these experiments the implementation of C4.5 algorithm was outperformed by other learning algorithms available in WEKA, i.e. Rotation Forest, Classification-

ViaRegression and Random Forest, however, it delivered better results than PART and Logistic.

Integrating the clustering algorithm (k-means) with decision trees (ID3) was also utilized to predict breast cancer (Khan, Mohamudally, 2011, p. 76–82). In the first step of the procedure authors divided the set of independent variables into 5 groups and on this basis they created 5 data sets consisting of the same number of observations and the same dependent variable. Then the k-means algorithm was applied to each data set separately with the fixed number of clusters that was equal to 5. Finally, decision tree models were built in the clusters producing different sets of "if...then..." rules. In the summary the authors attempted to visualize the outcomes, however, they did not compare the integrated model with the other data mining algorithms and did not use any performance measure.

## 3. Hybridization and datasets

### 3.1. Hybrid k-means + CART model

Authors call their approach "hybrid" since they use a sequential combination of unsupervised and supervised methods. Another reason for naming this approach "hybrid" is a combination of classical statistical tools (k-means method) with the algorithm derived from data mining (C&RT). In the first stage objects were clustered by using the k-means algorithm. In the second stage C&RT algorithm was applied, treating cluster membership of the objects as a new independent variable.

As the experiment involved the application of eight different datasets, the authors made an attempt to unify the procedure. It was decided that the set of variables utilized during the analysis of clusters will refer exclusively to numerical variables. The new categorical variable informing about the class membership was then attached to the remaining categorical variables, and the set completed in such a way constituted the basis for building a decision tree.

Data mining, apart from psychology, biology, statistics and machine learning, constitutes one of the most important areas in which the methods of cluster analysis are widely applied. Different variants of the k-means method result from the manner in which the initial positions of centroids are determined, the way of calculating centroids in successive steps of the algorithm, or the implemented measure of distance. In this work authors applied the Hartigan and Wong method (1979, p. 100–108), available in the R package `stats`.

A characteristic feature of the methods optimizing the initial partition of objects is determining *a priori* the number of clusters. One way to proceed in this area is establishing this number on the basis of classification quality measures. However, as emphasized by Everit et al. (Everit, Landau, Leese, Stahl, 2001), the selection of the optimal number of clusters should result from the synthesis of results obtained with the help of different methods. Such a conduct is justified by e.g. the fact that each of the methods is based on predefined assumptions referring to the class structure, which does not always have to be satisfied. Therefore, in this analysis we applied several measures frequently implemented in empirical research studies and available in the R package `clusterSim`:

- the Calinski-Harabasz index (CH) (Caliński, Harabasz, 1974, p. 1–27):

$$CH(k) = \frac{trace(W_k)/(k-1)}{trace(B_k)/(n-k)} \tag{1}$$

where $W_k$ and $B_k$ denote respectively within-group and between-group dispersion matrices. The optimal number of classes is indicated by the highest value of the index $CH(k)$.

- the Krzanowski-Lai index (KL) (Krzanowski, Lai, 1988, p. 23–34):

$$KL(k) = \left|\frac{DIFF_K}{DIFF_{k+1}}\right| \tag{2}$$

where: $DIFF_k = (k-1)^{2/p} trace(W_{k-1}) - u^{2/p} trace(W_k)$. The optimal number of clusters is indicated by the highest value of $KL(k)$.

- the Davies and Bouldin index (DB) (Davies, Bouldin, 1979, p. 224–227):

$$DB(k) = \frac{1}{k}\sum_{j=1}^{k} \max\left(\frac{c_j + c_l}{d_{jl}}\right) \tag{3}$$

where $C_j$ is the measure of the dispersion of objects in the *j*th cluster, and $d_{jl}$ is the distance between the centroids of clusters $j$ and $l$. The smallest $DB(k)$ indicates the optimal partition.

- the Hartigan index (H) (Hartigan, 1975):

$$H(k) = (n-k-1)\left(\frac{trace(W_k)}{trace(W_{k+1})} - 1\right) \tag{4}$$

In accordance with the Hartigan approach the optimal number of classes is $k$, which satisfies the condition $H(k) < \alpha$, where most frequently $\alpha = 10$ is assumed.

- the Gap Statistic (Gap) (Tibshirani, Walther, Hastie, 2001, p. 411–423):

$$Gap(k) = \frac{1}{B}\sum_{b=1}^{B} \log W_{kb} - \log W_k \tag{5}$$

where $B$ represents the number of the generated sets of objects. The optimal number of classes is the smallest value $k$ satisfying the condition $Gap(k) \geq Gap(k+1) - S_{k+1}$ where $S_{k+1}$ constitutes a factor that takes into account the standard deviation of the Monte-Carlo replicates $W_{kb}$.

Classification and Regression Trees (CART), which was developed by Breiman et al (Breiman, Friedman, Olshen, Stone, 1984), is a recursive partitioning algorithm. It is used to build a classification tree if the dependent variable is nominal, and a regression tree if the dependent variable is continuous. Decision trees usually do not have high predictive power. However, they deliver a set of rules and a graphical model that can be helpful in understanding the problem. The experiment involved the application of the C&RT algorithm with equal a priori probabilities and equal misclassification costs. The minimal number of instances in terminal nodes was established at the level of 5% of the learning sample.

### 3.2. Datasets used in experiment

The authors did their best to ensure that the datasets applied in the experiment refer to the marketing activity of companies. For this purpose they utilized popular repositories selecting datasets with a binary target variable. The first dataset refers to direct marketing campaigns of a Portuguese banking institution (subscribing a term deposit) (Moro, Laureano, Cortez, 2011, p. 117–121). The dependent variable in the second dataset (German Credit) is related to good or bad credit risks (Frank, Asuncion, 2010). The third dataset was used in the CoIL 2000 Challenge (van der Putten, van Someren, 2000). It is related to predicting the willingness to purchase a caravan insurance policy. The fourth dataset refers to direct marketing (response to mailing) and was used during KDD Cup 1998. The file is hosted on http://kdd.ics.uci.edu by I. Parsa and K. Howes. The fifth dataset includes target variable "churn" (Blake, Merz, 1998). The sixth dataset is also related to churn modeling and was used during KDD Cup in 2009 (http://www.kddcup-orange.com). The seventh dataset (CINA) consists of census data (http://www.causality.inf.ethz.ch/data/CINA.html).

The binary dependent variable indicates whether the income exceeds 50,000. The last dataset refers to credit card applications (Frank, Asuncion, 2010) with the binary target variable (approval/disapproval). The characteristics of all datasets, including size, number and kind of independent variables as well as the percentage of category "1" of the dependent variable was illustrated in Table 1.

**Table 1**

**Characteristics of datasets applied in experiment**

| Dataset | | Number of cases | Number of independent variables | Percentage of category "1" of dependent variable |
|---|---|---|---|---|
| D1 | Bank Marketing Data Set | 45,211 | 7 numerical and 9 categorical | 11.70% |
| D2 | Statlog (German Credit) | 1,000 | 7 numerical and 12 categorical | 30.00% |
| D3 | Insurance Company Benchmark | 5,822 | 80 numerical and 5 categorical | 5.98% |
| D4 | KDD 1998 | 95,412 | 286 numerical and 187 categorical | 5.08% |
| D5 | Churn | 5,000 | 16 numerical and 3 categorical | 14.14% |
| D6 | KDD 2009 | 50,000 | 190 numerical and 39 categorical | 7.34% |
| D7 | CINA Marketing Data Set | 16,033 | 21 numerical and 111 binary | 24.57% |
| D8 | Statlog (Australian Credit) | 690 | 6 numerical and 8 categorical | 44.49% |

Source: own research

Each set of observations was divided into the learning sample (70%) and the test sample (30%). In order to make the cluster interpretation simpler the number of variables applied while clustering could not exceed 15 (Blattberg, Kim, Neslin, 2008). If the dataset consisted of a larger number, a feature selection was undertaken with the help of Random Forests. The authors selected the ones with the highest variable importance score from the ranking of predictors. The variables for which the amount of missing data exceeded 10% as well as the cases for which the missing data exceeded 50% were removed from the sets. Eight categorical independent variables had a large number of categories (even exceeding 4000), which made it impossible to introduce dummy variables. The authors replaced them by an additional variable grouping them with the help of the EM algorithm into 12 categories. In cases where data were missing mean or mode were applied instead. The variables referring to ID, phone numbers and dates were excluded from the analysis.

## 4. Results of experiment

It was decided before starting the clustering procedure that the number of clusters cannot be larger than 15. The maximum number of clusters was determined by the CART algorithm requirements. It is assumed that independent variables should have fewer than 15 categories. Table 2 illustrates the optimal number of subgroups which was indicated by particular cluster validity measures. Lines (–) mean that in the range from 2 to 15 clusters no optimal number of classes was indicated by the measurement. It seems that the Davies-Bouldin index has a tendency to differentiate the highest number of clusters. On the other hand, the Hartigan index indicated the smallest number of subgroups or could not find an optimal solution at all. Hence, eventually 5 hybrid models were built on the basis of the eight datasets.

**Table 2**

**Number of clusters indicated by particular cluster validity measures**

| Dataset | Cluster validity measures | | | | |
|---|---|---|---|---|---|
| | CH | KL | DB | H | Gap |
| D1 | 6 | 6 | 9 | 2 | 3 |
| D2 | 15 | 11 | 7 | – | 6 |
| D3 | 2 | 15 | 2 | – | – |
| D4 | 2 | 6 | 12 | 2 | 2 |
| D5 | 2 | 5 | 12 | 2 | 15 |
| D6 | 4 | 12 | 15 | 4 | – |
| D7 | 2 | 4 | 14 | 2 | 4 |
| D8 | 5 | 8 | 11 | – | 2 |

CH – the Calinski-Harabasz index; KL – the Krzanowski-Lai index; DB – the Davies-Bouldin index; H – the Hartigan index; Gap – the gap statistic
Source: own research

The following popular performance measures were utilized for the assessment of models: accuracy ((TP+TN)/(TP+FP+TN+FN)), recall (TP/(TP+FN)), precision (TP/(TP+FP)), G-mean ((true negative rate x recall)$^{1/2}$), F-measure $(\frac{2 \times precision \times recall}{precision + recall})$, and lift in the first and in the second decile. The successive tables (3–8) contain the results for the eight datasets taken into account in the experiment as well as for 6 models. Five out of six decision tree models were modified by adding new categorical variables, while the sixth model remained unmodified. It was built on the basis of the entire set of independent variables (categorical and numerical). The table boxes highlighted with bold type signify that the hybrid model

**Table 3**

**Values of accuracy**

| Dataset | Hybrid CH | Hybrid KL | Hybrid DB | Hybrid H | Hybrid Gap | Unmodified decision tree model |
|---|---|---|---|---|---|---|
| D1 | 0.602 | 0.602 | **0.791** | no tree | no tree | 0.770 |
| D2 | 0.723 | **0.733** | **0.733** | – | **0.733** | 0.726 |
| D3 | **0.639** | 0.528 | **0.639** | – | – | 0.576 |
| D4 | **0.586** | **0.586** | **0.586** | **0.586** | **0.586** | 0.395 |
| D5 | 0.589 | 0.709 | 0.701 | 0.589 | 0.743 | 0.833 |
| D6 | 0.707 | 0.707 | 0.455 | 0.707 | – | 0.707 |
| D7 | 0.881 | 0.897 | **0.907** | 0.881 | 0.897 | 0.905 |
| D8 | 0.851 | 0.851 | 0.851 | – | 0.851 | 0.862 |

Source: own research

reached a higher value of the quality measurement than the unmodified model.

If the values of accuracy (Table 3) are to be taken into account, one can clearly see that the best hybrid models were created with the number of clusters indicated by the Davies-Bouldin index (DB). “No tree” means that with a given set of independent variables and parameters of the algorithm no tree was grown. The structure of the tree would require a modification of a priori probabilities or misclassification costs, which the authors desired to avoid in order to maintain the standard procedure.

**Table 4**

**Values of recall**

| Dataset | Hybrid CH | Hybrid KL | Hybrid DB | Hybrid H | Hybrid Gap | Unmodified decision tree model |
|---|---|---|---|---|---|---|
| D1 | 0.512 | 0.512 | 0.437 | no tree | no tree | 0.712 |
| D2 | **0.495** | **0.680** | **0.680** | – | **0.680** | 0.454 |
| D3 | 0.579 | 0.738 | 0.579 | – | – | 0.813 |
| D4 | 0.545 | 0.545 | 0.545 | 0.545 | 0.545 | 0.725 |
| D5 | 0.636 | 0.636 | 0.790 | 0.636 | 0.785 | 0.827 |
| D6 | 0.375 | 0.375 | **0.674** | 0.375 | – | 0.375 |
| D7 | **0.907** | **0.903** | **0.901** | **0.907** | **0.903** | 0.885 |
| D8 | 0.843 | 0.843 | 0.843 | – | 0.843 | 0.892 |

Source: own research

In the case of recall values (Table 4) the results for hybrid models were better in datasets: D2, D6, and D7. It is hard to indicate which cluster validity measure is the best. However, it seems that Davies-Bouldin index delivers higher values of recall more often than other indices.

**Table 5**

**Values of precision**

| Dataset | Hybrid CH | Hybrid KL | Hybrid DB | Hybrid H | Hybrid Gap | Unmodified decision tree model |
|---|---|---|---|---|---|---|
| D1 | 0.147 | 0.147 | 0.259 | no tree | no tree | 0.294 |
| D2 | 0.571 | 0.564 | 0.564 | – | 0.564 | 0.587 |
| D3 | 0.095 | 0.090 | 0.095 | – | – | 0.107 |
| D4 | **0.066** | **0.066** | **0.066** | **0.066** | **0.066** | 0.058 |
| D5 | 0.199 | 0.272 | 0.292 | 0.199 | 0.328 | 0.449 |
| D6 | 0.098 | 0.098 | 0.085 | 0.098 | – | 0.098 |
| D7 | 0.699 | 0.740 | 0.764 | 0.699 | 0.740 | 0.766 |
| D8 | **0.824** | **0.824** | **0.824** | – | **0.824** | 0.813 |

Source: own research

As far as precision is concerned (Table 5), the best results were achieved in datasets D4 and D8. However, more often solutions were identical with the ones in the unmodified model (dataset D6) or worse (D1, D2, D3, D5, D7).

**Table 6**

**Values of G-mean**

| Dataset | Hybrid CH | Hybrid KL | Hybrid DB | Hybrid H | Hybrid Gap | Unmodified decision tree model |
|---|---|---|---|---|---|---|
| D1 | 0.561 | 0.561 | 0.605 | no tree | no tree | 0.744 |
| D2 | **0.640** | **0.718** | **0.718** | – | **0.718** | 0.622 |
| D3 | 0.610 | 0.616 | 0.610 | – | – | 0.675 |
| D4 | **0.566** | **0.566** | **0.566** | **0.566** | **0.566** | 0.523 |
| D5 | 0.608 | 0.677 | 0.736 | 0.608 | 0.760 | 0.830 |
| D6 | 0.524 | 0.524 | **0.543** | 0.524 | – | 0.524 |
| D7 | 0.890 | **0.899** | **0.905** | 0.890 | **0.899** | 0.898 |
| D8 | 0.850 | 0.850 | 0.850 | – | 0.850 | 0.864 |

Source: own research

Considering the values of G-mean (Table 6) one may observe a relatively large effectiveness of hybrid models, in particular those based on the Davies-Bouldin index (DB), Krzanowski-Lai index (KL) and the gap statistic (Gap).

**Table 7**

**Values of F-measure**

| Dataset | Hybrid CH | Hybrid KL | Hybrid DB | Hybrid H | Hybrid Gap | Unmodified decision tree model |
|---|---|---|---|---|---|---|
| D1 | 0.228 | 0.228 | 0.325 | no tree | no tree | 0.416 |
| D2 | **0.530** | **0.617** | **0.617** | – | **0.617** | 0.512 |
| D3 | 0.164 | 0.160 | 0.164 | – | – | 0.190 |
| D4 | **0.118** | **0.118** | **0.118** | **0.118** | **0.118** | 0.108 |
| D5 | 0.303 | 0.381 | 0.426 | 0.303 | 0.463 | 0.582 |
| D6 | 0.155 | 0.155 | 0.151 | 0.155 | – | 0.155 |
| D7 | 0.790 | 0.813 | **0.827** | 0.790 | 0.813 | 0.821 |
| D8 | 0.833 | 0.833 | 0.833 | – | 0.833 | 0.851 |

Source: own research

As far as the F-measure is concerned (Table 7) one can again see the advantage of the Davies-Bouldin index (DB). Hybrid models proved to be better in datasets: D2, D4 and D7.

The authors' anticipations regarding the values of the lift measure (Table 8) were not fully confirmed. In four out of eight datasets hybrid models outperformed the unmodified decision tree model. It concerns datasets: D2, D7, and D8 in both deciles and dataset D6 in the first decile. It is worth noting that the results are better than in the experiment, in which the minimum number of cases in terminal nodes was established at the level of 10% of the learning sample (Łapczyński, Jefmański, 2013).

**Table 8**

**Values of lift in 1st decile and in 2nd decile**

| Dataset | Hybrid CH | | Hybrid KL | | Hybrid DB | | Hybrid H | | Hybrid Gap | | Unmodified decision tree model | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | 1st | 2nd | 1st | 2nd | 1st | 2nd | 1st | 2nd | 1st | 2nd | 1st | 2nd |
| D1 | 1.2 | 1.2 | 1.2 | 1.2 | 2.3 | 1.3 | no tree | | no tree | | 3.6 | 2.6 |
| D2 | **2.1** | 1.9 | **2.0** | 1.8 | **2.0** | **2.0** | – | – | **2.0** | **2.0** | 1.8 | 1.9 |
| D3 | 2.0 | 1.6 | 2.3 | 1.7 | 2.0 | 1.6 | – | – | – | – | 2.5 | 2.0 |
| D4 | 1.4 | 1.4 | 1.4 | 1.4 | 1.4 | 1.4 | 1.4 | 1.4 | 1.4 | 1.4 | 1.6 | 1.6 |
| D5 | 3.0 | 1.9 | 2.6 | 2.6 | 2.8 | 2.8 | 3.0 | 1.9 | 2.7 | 2.7 | 3.5 | 3.2 |
| D6 | **1.6** | 1.4 | **1.6** | 1.4 | 1.3 | 1.3 | **1.6** | 1.4 | – | – | 1.4 | 1.4 |
| D7 | **3.5** | **3.5** | **3.5** | **3.5** | **3.6** | **3.6** | **3.5** | **3.5** | **3.5** | **3.5** | 3.1 | 3.1 |
| D8 | **2.1** | **2.1** | **2.1** | **2.1** | **2.3** | **2.2** | – | – | **2.1** | **2.1** | 1.8 | 2.0 |

Source: own research

The differences between the performance of hybrid models may be caused by various factors. One of them is a result of a large number of combinations of categorical predictors divisions. The number of possible partitions of the nominal variable, e.g. cluster membership is equal to $2^{n-1} - 1$ (where $n$ represents the number of categories of that variable) while the number of possible partitions for numerical predictors is equal to $n$ (the number of the value of the quantitative variable). A larger number of potential splits of the node increases the probability of finding the optimal solution.

On the other hand, a low performance of hybrid models may result from the double application of the test sample. In the first step cases are assigned to clusters, whereas in the successive step the predictive model is implemented.

Sets D1 and D4 are characterized by the highest skewness of the distribution of variables forming clusters and by the highest outliers values. Some of them possess standardized values exceeding 100 with the standard deviation equal to 1. Skewness measures are very high – considerably exceeding the value 0 (e.g. 113, 48, 29). This may lead to a low quality of hybrid models. Similar distributions of quantitative variables characterized the set D1; however it was possible to obtain satisfying results of accuracy and precision there.

A low quality of hybrid models may also be caused by the lack of 'natural clusters' in the dataset. The algorithm may have provided artefactual solutions. A potential cause of the failure may be determined by the use of the Euclidean distance instead of Mahalanobis distance. After the experiment it appeared that in some of the data sets (D1, D6, D8) the nonspherical shape of the clusters is present.

Finally, one can assume that the failure is due to a low impact of quantitative independent variables on the dependent variable in non-modified models. The variable importance rankings lead to the conclusion that these variables were not important in predicting the dependent variable.

## 5. Conclusions

The construction of hybrid models based on the k-means algorithm and C&RT decision trees may in some situations improve the performance of predictive models. It appears that cluster validity indices, which determine a different optimal number of clusters, play an important role here. It may be concluded from the conducted experiment that the Davies-Bouldin index proves to perform the best. Hybrid models supply higher values of accuracy and G-mean. In some cases they are better as far as F-measure, recall and precision are concerned. It also seems that they deliver promising results when it comes to improving the lift measure, which plays an important role in marketing application.

The best results are obtained in the case of hybrid models, in which the number of clusters is relatively high. This constitutes a certain inconvenience as an excessively high number of subgroups complicates their interpretation. No connection was noted between the performance measures of hybrid models and the percentage of class "1" of the dependent variable.

The authors see the need for the extension of the experiment onto other datasets, the modification of parameters of the decision tree (e.g. a priori probabilities, misclassification costs and minimum number of instances in the terminal node), the differentiation of distances (Euclidean or Mahalanobis), elimination of outliers, and experiments with fuzzy clustering methods.

## NOTES

[1] The project was financed by a grant from National Science Centre (DEC-2011/01/B/HS4/04758)

## REFERENCES

Blake, C.L., Merz, C.J. (1998) Churn Data Set, UCI Repository of Machine Learning Databases. http://www.sgi.com/tech/mlc/db, University of California, Department of Information and Computer Science, Irvine, CA.

Blattberg, R.C., Kim, B-D, Neslin, S.A., (2008) *Database Marketing. Analyzing and Managing Customers*, New York: Springer.

Bose, I., Chen, X. (2009). Hybrid Models Using Unsupervised Clustering for Prediction of Customer Churn. *Journal of Organizational Computing and Electronic Commerce*. vol. 19, no. 2, April-June, 133–151.

Breiman, L., Friedman, J.H., Olshen, R.A., Stone, C.J. (1984). *Classification and Regression Trees*. Belmont, CA: Wadsworth International Group.

Caliński, R.B, Harabasz, J. (1974). A Dendrite Method for Cluster Analysis. *Communications in Statistics*. vol. 3, iss. 1, 1–27.

Causality Workbench. Challenges in Machine Learning, http://www.causality.inf.ethz.ch/data/CINA.html.

Christopher, M., Payne, A., Ballantyne, D. (2002). *Relationship Marketing. Creating Stakeholder Value*. Oxford: Elsevier.

Chu, B-H., Tsai, M-S., Ho, Ch-S. (2007). Toward a Hybrid Data Mining Model for Customer Retention. *Knowledge-Based Systems.* no. 20, 703–718.

Davies, D.L., Bouldin, D.W. (1979). A Cluster Separation Measure. In *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 1, no. 2, 224–227.

Everit, B.S., Landau, S., Leese, M., Stahl, D. (2011). *Cluster Analysis. 5th Edition.* Chichester: John Wiley & Sons.

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 D. Ryzko et al. *Emerging Intelligent Technologies in Industry*, SCI 369, Berlin Heidelberg: Springer-Verlag, 21–34.

Frank, A., Asuncion, A. (2010). UCI Machine Learning Repository [http://archive.ics.uci.edu/ml]. Irvine, CA: University of California, School of Information and Computer Science.

Gaddam, S.R., Phoha, V.V., Balagani, K.S. (2007). K-means + ID3: A Novel Method for Supervised Anomaly Detection by Cascading K-means Clustering and ID3 Decision Tree Learning Methods. In: *IEEE Transactions on Knowledge and Data Engineering*, vol. 19, no. 3, March, 345–354.

Hartigan, J.A. (1975). *Clustering Algorithms.* New York, London, Sydney, Toronto: Wiley.

Hartigan, J.A., Wong, M.A. (1979). A K-means Clustering Algorithm. *Applied Statistics.* vol. 28, no. 1, 100–108.

KDD Cup 2009, http://www.kddcup-orange.com.

Khan, D.M., Mohamudally, N. (2011). An Integration of K-means and Decision Tree (ID3) Towards a More Efficient Data Mining Algorithm. *Journal of Computing.* vol. 3, iss. 12, December, 76–82.

Krzanowski, W.J., Lai, Y.T. (1988). A Criterion for Determining the Number of Groups in a Data Set Using Sum-of-Squares Clustering. *Biometrics.* vol. 44, no. 1, 23–34.

Kumar, V., Rathee, N. (2011). Knowledge Discovery from Database Using an Integration of Clustering and Classification. *International Journal of Advanced Computer Science and Applications.* vol. 2, no. 3, March, 29–33.

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

Łapczyński, M., Surma, J. (2012). Hybrid Predictive Models for Optimizing Marketing Banner Ad Campaign in On-line Social Network. In R. Stahlbock, G.M. Weiss (Eds.) *Proceedings of the 2012 International Conference on Data Mining*, Las Vegas Nevada, USA: CSREA Press, 140–146.

Li, Y., Deng, Z., Qian, Q., Xu, R. (2011). Churn Forecast Based on Two-step Classification in Security Industry. *Intelligent Information Management.* no. 3, 160–165.

Moro, S., Laureano, R., Cortez, P. (2011). Using Data Mining for Bank Direct Marketing: An Application of the CRISP-DM Methodology. In P. Novais et al. (Eds.) *Proceedings of the European Simulation and Modelling Conference – ESM'2011*, Guimarães, Portugal, October, 117–121.

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 R. Stahlbock, G.M. Weiss (Eds.) *Proceedings of the 2012 International al Conference on Data Mining*, Las Vegas Nevada, USA: CSREA Press, 24–30.

Tibshirani, R., Walther, G., Hastie, T. (2001). Estimating the Number of Clusters in a Data Set via the Gap Statistic. *Journal of the Royal Statistical Society.* ser. B, 63, part 2, 411–423.

van der Putten, P., van Someren, M. (Eds) (2000). CoIL Challenge 2000: The Insurance Company Case. In Also a Leiden Institute of Advanced Computer Science Technical Report 2000–09, Sentient Machine Research, Amsterdam, June 22.

Wierenga, B. (Ed.) (2008). *Handbook of Marketing Decision Models.* New York: Springer.