8-2010

A method for business sequential data prediction

Tomasz Kajdanowicz
Wroclaw University of Technology, tomasz.kajdanowicz@pwr.wroc.pl

Przemysław Kazienko
Wroclaw University of Technology, kazienko@pwr.wroc.pl

A method for business sequential data prediction

Tomasz Kajdanowicz
Wroclaw University of Technology
Wyb. Wyspianskiego 27, 50-370 Wroclaw, Poland
tomasz.kajdanowicz@pwr.wroc.pl

Przemysław Kazienko
Wroclaw University of Technology
Wyb. Wyspianskiego 27, 50-370 Wroclaw, Poland
kazienko@pwr.wroc.pl

ABSTRACT

The paper describes the process and new methodology of a hybrid prediction system for business sequential data i.e. debt portfolio appraisal. Conducting the local area competence data mining approach, repayment values are predicted by means of combination of various data mining techniques. The methods include clustering of references, model selection and enrichment of input variables with prediction outputs from preceding periods. Experimental studies concern the method's configuration influence on its general performance such as number of distinct predictors and number of competence areas.

Keywords

Claim appraisal, competence area modeling, decision support system, financial prediction, repayment prediction

INTRODUCTION

Debt portfolio valuation is a prediction task that assesses the repayment value from the debt cases. Business models, which relay on the cash flow from receivables, assume to keep the aggregated debt value as small as it is possible. The fact that possession of creditors for a long time is very ineffective, especially when debtors are eventually not able to repay theirs arrears in short term, implies a need of sophisticated debts valuation. Under some circumstances, it is better for companies to sell the liabilities to a specialized debt collection company in order to obtain at least a part of their nominal value, rather than collect and vindicate debts on their own. In the process of selling a debt portfolio the transaction price is usually estimated based on the possible repayment level to be reached in the long term. In general, it is expected that the method of debt portfolio value appraisal will well match the future recovery process.

RELATED WORK

Prediction and classification has been widely studied and described in the literature, e.g. [2, 3, 12, 16]. Overall, the existing machine learning methods usually provide better classification and prediction accuracy than techniques based on common statistical techniques such as regression [17, 18]. However, a better precision of the prediction may be obtained by combination of several existing methods into one hybrid solution [1, 2, 4]. In general, hybridization could be achieved either by application of additional external mechanisms into existing prediction models (low level), e.g. neuro-fuzzy systems [11] or by combination of different methods on the high level, e.g. multiple classifier systems, where separate classifiers are treated more likewise 'black boxes' [6, 7]. Hybrid prediction methods have been successfully used in a number of domains such as medicine, engineering and industry. Other application areas of these methods are economy and finance, where hybrid systems provide specialized knowledge in order to support business decisions [15].

The paper is focused on the description of a new hybrid method for debt portfolio appraisal. The correct prediction of target repayment value in debt recovery is of great practical importance, because it reveals the level of possible expected benefit and chances to collect receivables. The crucial concept of this method is the combination of clustering of the training set and application of multiple classifiers based on their competence regions [12]. Additionally, a sequence of classifiers is built to obtain predictions over consecutive periods. Apart from the general idea, the proposed hybrid prediction method has been examined on real data. According to the findings achieved, the method appears to return more precise results compared to some common approaches.

CLAIM APPRAISAL

The process of debt portfolio value prediction starts when the first company offers a package of debts and expects a purchase proposal from the second one. The second company is usually a specialized debt recovery entity. Based on historical data of debt recovery available for the second company, a prediction model is prepared. The model provides estimation of possible

return from the package. The bid is supplemented by additional cost of repayment procedures and cash flow abilities as far as risk and final purchase price are proposed to the first company. The most significant and sensitive part of the process is the repayment value prediction for debt portfolio as there is a strong business need for the method to be designed for the efficient and accurate prediction, which will respect the time factor.

Having the data of historical claim cases together with their repayment profiles over time, a debt collection company can build a model in order to predict receivables for the new claim set invited for bids. However, in order to be able to evaluate cash flows in the following periods (usually months), the company needs to have a possibly precise distribution of the receivables collection. It helps to estimate the final upper value for the considered input debt portfolio. Hence, not only the total aggregated value of the receivables is useful for bidding but also their probable timing, period by period.

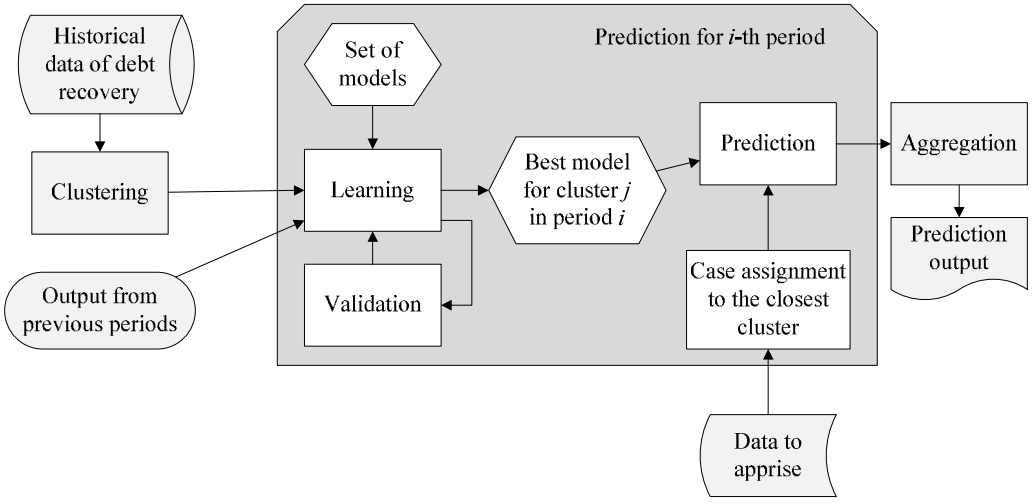

Figure 1. Concept of sequential prediction of sequential debt portfolio valuation.

The idea of the hybrid method for prediction of debt recovery value consists of data flows that are executed separately for each period i (M times), Figure 1. First, the prepared historical data is clustered into groups of similar debt cases – competence regions [12, 14]. Next, a set of models is created (learnt) separately for each cluster j using the fixed set of common, predefined models. The best one is selected for each cluster and becomes the cluster's predictive model P_{ij} for period i. This assignment is done based on minimization of the standard deviation error. This main learning phase is followed by the final prediction for the debt portfolio. For each of debt case, the closest cluster of historical data is determined and the prediction for this case is performed based on the model trained and assigned to that cluster, separately for each period i.

The important characteristic of the method is that the predicted value of return on debt in period i is taken as the input variable for the i+1th period prediction as an additional feature in the model, see Figure 2.

Historical, reference cases are in general clustered into NG groups using partitioning method and the best prediction model P_{ij} is separately assigned to each group j and each period i. Features directly available within the input data set or new ones derived from them are the only used in the clustering process. Besides, clustering is performed for the whole reference set, i.e. for cases being after at least one period of the recovery procedure (period 1). For the following periods, e.g. for period i, cases with too short history (being recovered shorter than i periods), are just removed from their clusters without re-clustering. As a result, the quantity of each cluster G_{ij} may vary depending on the period i and it is smaller for greater i. For the ith period and the jth group G_{ij}, we have: $\text{card}(G_{ij}) \geq \text{card}(G(i+1)j)$. In consequence, there are the same reference groups for all periods but their content usually decreases for the following periods. This is obvious, because the debt collection company possesses many pending recovery cases, which can be used as references in prediction only for the beginning periods. If the quantity of one cluster for the greater period is too small then this cluster is merged with another, closest one for all following periods.

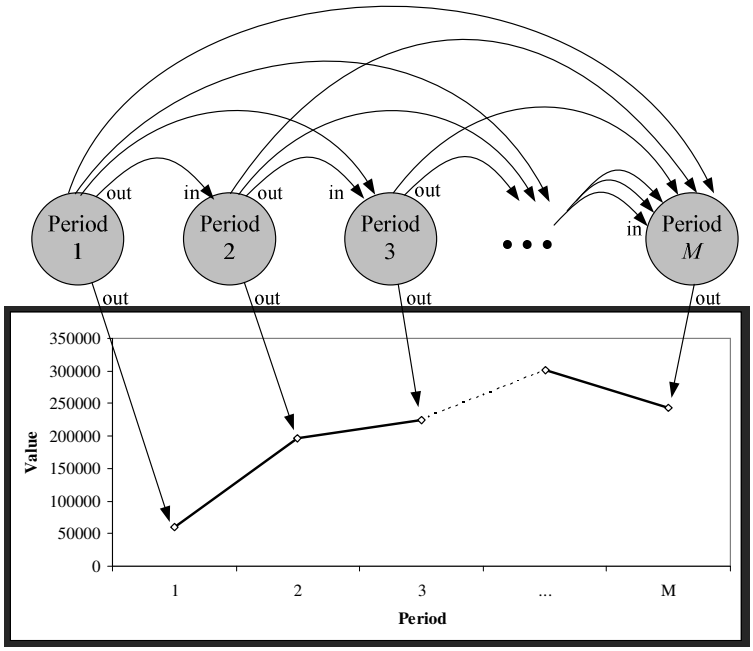

Figure 2. Input variable dependency in sequential prediction of debt repayment

Each group G_{ij} possesses its own, separate representation and the common similarity function is used to evaluate closeness between group G_{ij} and each input case x just being predicted. Next, the single closest group, or more precise the assigned model, is applied to the input case x. Afterwards, predicted values are aggregated from all input cases. As a result, we obtain a sequence of consecutive values, as in Figure 2.

ENVIRONMENT SETUP

For the experimental examination of the proposed method 12 distinct real debt recovery data sets were used. A summary of the data profile is presented in Table 1. In total, 20 input features were extracted: 5 continuous, 9 nominal and 6 binary. The experiments were performed using 5 cross-fold validation setup independently applied for each data set [5]. Three distinct experiments were conducted: 1) examination of the relationship between the hybrid complexity and the prediction accuracy, 2) the influence of the number of groups (competence regions) on the prediction accuracy and 3) the impact of predictors' diversity on the accuracy. Seven predictors were used in the experiments, see Table 2. The best predictor assignment to each group is carried out based on the minimization of the prediction error standard deviation. The research was implemented and conducted within the R statistical computing environment with some extended and customized algorithms based on RWeka, rJava and tree plug-ins. In the experiment, the debt recovery value prediction has been conducted for 10 consecutive periods (months).

Data set	No. of cases	Data set	No. of cases
A	4019	G	6818
B	3440	H	1703
C	2764	I	4584
D	3175	J	6607
E	3736	K	2515
F	4211	L	1104

Table 1. Summary of debt recovery data sets

No.	Name	Settings	Description
1	M5P	pruned M=4	Pruned regression tree M5P M – min. leaf capacity
2	M5P	pruned M=10	
3	Tree	F=0.2 N=4	Regression tree χ^2 F – spliting criterion N – min. leaf capacity
4	LineRegression	S=0	Line Regression M5 method for attribute selection
5	DecisionStump		One rule regression tree
6	Bagging	I=10 P=100 V=0.001	I – no. of iterations P – % size of subgroup V – min. variance for split
7	Bagging	I=10 P=100 V=0.01	

Table 2. Methods used in experiments

In the first experiment, three different scenarios were realized and finally compared with each other. In each scenario, the output of period i is used as the input variable for the following periods, Figure 2. The first scenario assumes simple prediction to be carried out on the single model (regression tree), which is learnt and validated on the training data without any clustering. The learning is accomplished separately for each period. The first scenario can be treated as the basic approach for value prediction of sequential and continuous variables. In the second scenario, also without clustering, the selection from seven distinct predictors is performed separately for each period. Only one, the best is chosen for each period. The full hybrid process is performed in the third scenario. It includes especially the clustering of the reference data set, see Figure 1 and 2. Clustered data was used to train all models and the best model was determined for each cluster and each period. Next, in the testing phase, the input cases were assigned to the closest cluster and processed by the assigned predictor. In other words, if an appraisal case is close to a certain cluster, the return value would be predicted by the model assigned to that cluster. The second scenario extends the first one, whereas the third expands the second.

The second experiment similarly assumes three scenarios. All of them are conducted on seven predictors but for the different number of groups.

The third experiment examines and compares the variability of the best predictors' participation in the overall prediction. The experiment parameters are presented in Table 3.

Experiment	Scenario	No. of predictors	Clustering groups	Generated predictors
1 hybridization vs. accuracy	1	1	1	50
	2	7	1	350
	3	7	5	1750
2 clustering vs. accuracy	1	7	1	350
	2	7	5	1750
	3	7	10	3500
3 clustering vs. diversity	1	7	1	350
	2	7	10	3500

Table 3. Experiment parameters

EXPERIMENTAL RESULTS

Having the methods for debt appraisal established, three experiments were launched. Each scenario from all the experiments has been compared with each other in respect of average prediction accuracy. The results of experiments are presented in Table 4, Table 5 and Table 6.

	Scenario 1 – one predictor			Scenario 2 – best predictor			Scenario 3 – clustering with best predictor		
Data set	Acc	Err	ErrV	Acc	Err	ErrV	Acc	Err	ErrV
A	0.50	0.39	0.21	0.73	0.21	0.03	0.84	0.28	0.03
B	0.53	0.46	0.07	0.83	0.20	0.01	0.75	0.19	0.01
C	0.69	0.31	0.04	0.89	0.13	0.01	0.82	0.31	0.01
D	0.80	0.40	0.11	0.88	0.13	0.01	0.89	0.19	0.01
E	0.73	0.30	0.03	0.92	0.09	0.00	0.83	0.20	0.01
F	0.68	0.32	0.04	0.89	0.16	0.02	0.87	0.27	0.04
G	0.97	0.18	0.03	0.97	0.02	0.00	0.96	0.07	0.00
H	0.82	0.22	0.02	0.96	0.07	0.00	0.97	0.15	0.01
I	0.44	0.33	0.17	0.82	0.22	0.03	0.68	0.39	0.03
J	0.71	0.23	0.12	0.97	0.06	0.00	0.93	0.10	0.00
K	0.69	0.32	0.07	0.95	0.07	0.00	0.84	0.18	0.02
L	0.83	0.19	0.11	0.87	0.11	0.01	0.86	0.20	0.02
Avg.	0.50	0.39	0.21	0.73	0.21	0.03	0.84	0.28	0.03

Acc = prediction accuracy, Err = prediction error, ErrV = prediction error variance

Table 4. The prediction accuracy results of experiment 1 for three different scenarios

Experiment 1 – influence of hybridization on prediction accuracy

The results of three distinct prediction scenarios revealed that the first scenario performs worse than the second and third by about 20%. The second and third methods for debt portfolio valuation are not significantly different in terms of prediction accuracy.

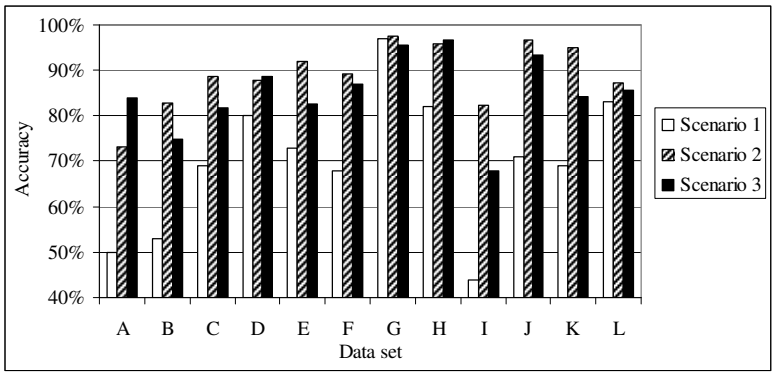

Figure 3. The influence of hybridization on accuracy of debt recovery value prediction for each data set from A to L

Studying ensemble like approaches, it is worth analyzing the error performance in terms of bias and variance factors. The bias and variance reflects the contribution of the prediction error for consecutive periods to the general error prediction [8]. Although, it may happen that the aggregated value of prediction for all periods reveals smaller error rate than the sum of errors from all periods, the prediction for the single period may overestimate or underestimate. Bias and variance of the prediction is presented in the Figure 4 and Figure 5, appropriately.

Studying ensemble like approaches, it is worth analyzing the error performance in terms of bias and variance factors. The bias and variance reflects the contribution of the prediction error for consecutive periods to the general error prediction [8]. Although, it may happen that the aggregated value of prediction for all periods reveals smaller error rate than the sum of errors from all periods, the prediction for the single period may overestimate or underestimate. Bias and variance of the prediction is presented in the Figure 4 and Figure 5, appropriately.

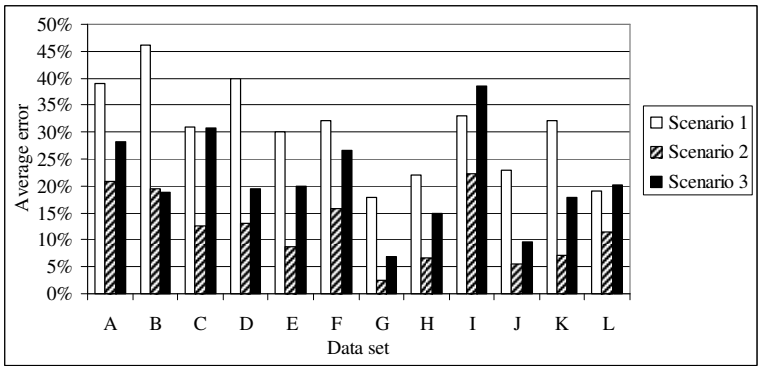

Figure 4. Bias composition for sequential (period) prediction error in experiment 1 (the influence of hybridization on accuracy)

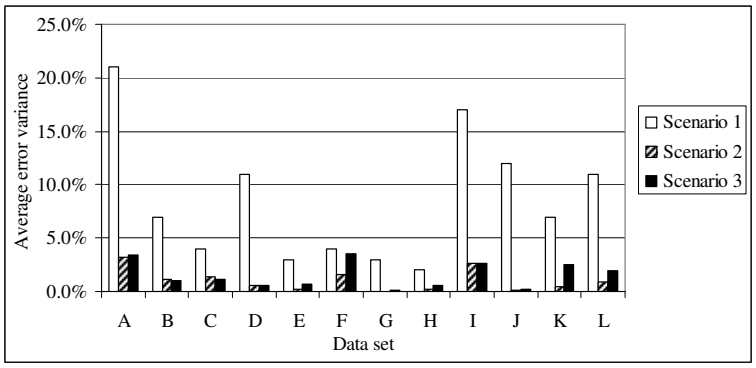

Figure 5. Variance composition for sequential (period) prediction error in experiment 1

As seen in Figure 4 and Figure 5, methods from the second and third scenario are more stable with respect to prediction error over time. That is directly reflected in low variance term, concerning most of the error in the bias.

Experiment 2 – influence of the number of clusters on prediction accuracy

The analysis of results from Table 5 shows that an increase of the number of clustering groups does not result much in the prediction accuracy improvement. In the first scenario, that assumes no clustering, average error on prediction reached the lowest value in comparison with the second and third scenario. All scenarios are similar in terms of error stability, see Figure 7 and Figure 8.

	Scenario 1 - one predictor			**Scenario 2 – best predictor**			**Scenario 3 – clustering with best predictor**		
Data set	Acc	Err	ErrV	Acc	Err	ErrV	Acc	Err	ErrV
A	0.73	0.21	0.03	0.84	0.28	0.03	0.81	0.26	0.02
B	0.83	0.20	0.01	0.75	0.19	0.01	0.89	0.19	0.00
C	0.89	0.13	0.01	0.82	0.31	0.01	0.85	0.27	0.02
D	0.88	0.13	0.01	0.89	0.19	0.01	0.91	0.23	0.01
E	0.92	0.09	0.00	0.83	0.20	0.01	0.84	0.25	0.02
F	0.89	0.16	0.02	0.87	0.27	0.04	0.95	0.24	0.00
G	0.97	0.02	0.00	0.96	0.07	0.00	0.95	0.10	0.00
H	0.96	0.07	0.00	0.97	0.15	0.01	0.97	0.17	0.01
I	0.82	0.22	0.03	0.68	0.39	0.03	0.72	0.42	0.03
J	0.97	0.06	0.00	0.93	0.10	0.00	0.95	0.11	0.00
K	0.95	0.07	0.00	0.84	0.18	0.02	0.84	0.19	0.02
L	0.87	0.11	0.01	0.86	0.20	0.02	0.91	0.19	0.01
Avg.	0.89	0.12	0.01	0.85	0.21	0.02	0.88	0.22	0.01

Acc = prediction accuracy, Err = prediction error, ErrV = prediction error variance

Table 5. The results of experiment 2 for three different scenarios – the influence of the number of clusters on prediction accuracy

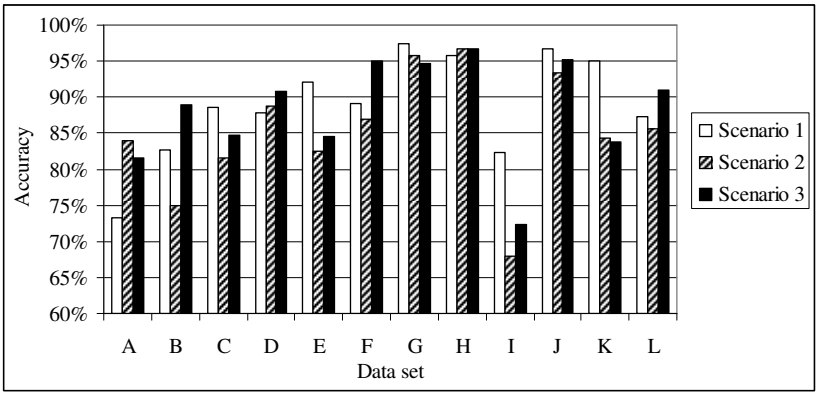

Figure 6. The influence of the number of clusters on prediction accuracy for data set from A to L

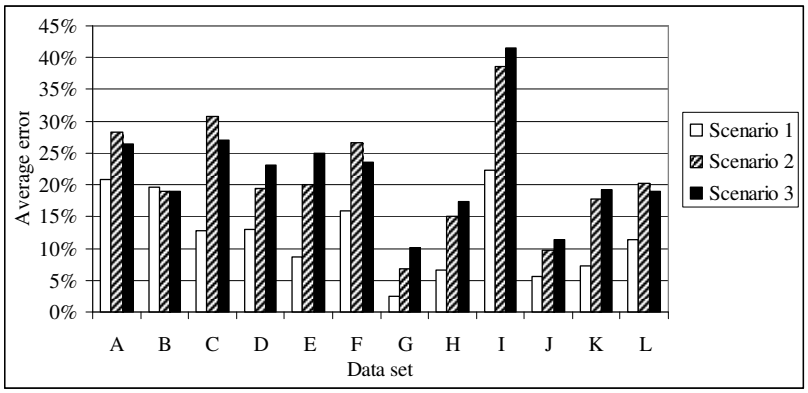

Figure 7. Bias composition for sequential (period) prediction error in experiment 2 (the influence of the number of clusters on prediction accuracy

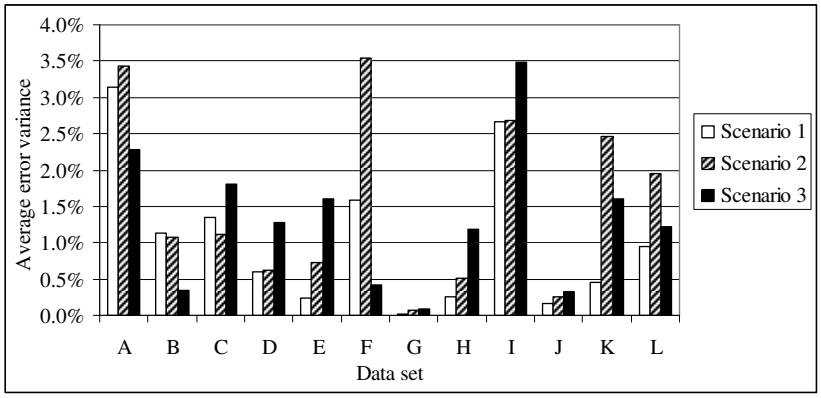

Figure 8. Variance composition for sequential (period) prediction error in experiment 2 (the influence of the number of clusters on prediction accuracy)

Experiment 3 – number of clusters vs. diversity

In terms of assessment of distinct predictors' participation in the overall prediction, the experiment was examined using two scenarios. The first one assumed no clustering, while the second one included clustering with 10 groups. In both cases, the variability of distinct predictors was followed, see Table 6, Figure 9 and Figure 10 for the results.

In scenario 1, the most often applied predictors were the most complex – bagging was predicting in 96% of cases. On one hand, this technique is relatively complex in terms of computational cost, in comparison to other predictive methods. However, it better reflects the structure of the data than the simpler methods. On the other hand, in the scenario 2, while the prediction is performed based on primarily computed clusters, the variety of applied classifiers is more diverse. From 10% to 50% of the utilized predictors are the simplest approaches like linear regression. It reveals that the clustering extracts the more coherent competence regions, i.e. it splits the complex structure of the data diversity into simpler small groups, in which simple machine learning concepts are performing quite efficiently.

	Models participation in prediction	
Predictor	**Scenario 1**	**Scenario 2**
M5P 1	0.03	0.09
M5P 2	0.01	0.06
Tree	0.00	0.03
LineRegression	0.00	0.03
DecisionStump	0.00	0.07
Bagging 1	0.62	0.41
Bagging 2	0.34	0.30

Table 6. The results of experiment 3 for two distinct scenarios

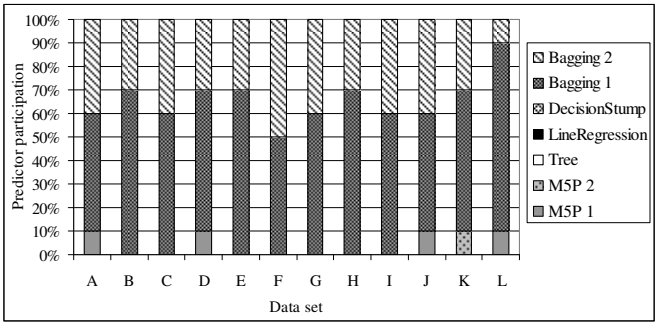

Figure 9. Models participation in prediction for scenario 1

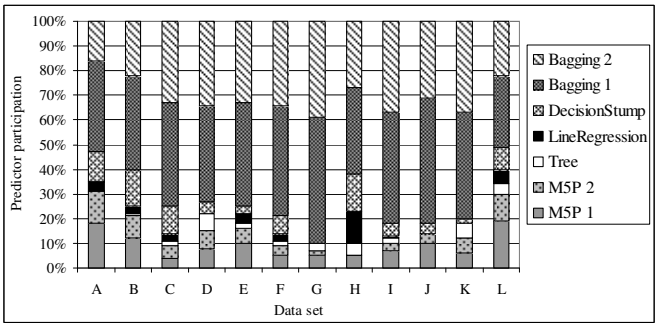

Figure 10. Models participation in prediction for scenario 2

CONCLUSIONS AND FUTURE WORK

In order to predict debt portfolio value, the proper hybrid method has been suggested and examined on real data. The experimental results support the conclusion that combined prediction solutions are performing well in terms of accuracy and stability as well as may be efficiently applied to debt recovery valuation.

In the future studies, many further aspects improving the proposed method will be considered; in particular: further combination of distinct types of classifiers, models' tuning using genetic based optimization [13] and adaptive clustering.

The application of the similar hybrid concept is also considered to be applied to social-based recommender systems [10].

ACKNOWLEDGMENTS

The work was supported by The Polish Ministry of Science and Higher Education, the development project, 2009-11.

REFERENCES

1. Aburto L., Weber R., A Sequential Hybrid Forecasting System for Demand Prediction, MLDM 2007, 2007, pp. 518-532.
2. Ali S., Smith K, On learning algorithm selection for classification, Applied Soft Computing, 6(2), 2006, pp. 119-138.
3. Bishop C. M., Pattern Recognition and Machine Learning, Springer, 2006.
4. Chou C.-H., Lin C.-C., Liu Y.-H., Chang F., A prototype classification method and its use in a hybrid solution for multiclass pattern recognition, Pattern Recognition, 39(4), 2006, pp. 624-634.
5. Dietterich T. G., Approximate statistical tests for comparing supervised classification learning algorithms, Neural Computation, 10(7), 1998, pp. 1895-1923.
6. Eastwood M., Gabrys B., Building Combined Classifiers, A chapter in Knowledge Processing and Reasoning for Information Society. Nguyen N.T., Kolaczek G., Gabrys B. (Eds.), EXIT Publishing House, Warsaw, 2008, pp. 139-163.
7. Gabrys B., Ruta D., Genetic algorithms in classifier fusion, Applied Soft Computing, 6(4) , 2006, pp. 337-347.
8. Garcia-Pedrajas N., Ortiz-Boyer D., Boosting k-nearest neighbor classifier by means of input space projection, Expert Systems with Applications, 36, 2009, pp. 10570-10582.
9. Kajdanowicz, T., Kazienko P., Hybrid Repayment Prediction for Debt Portfolio, ICCCI'09, LNAI 5796, 2009, pp. 850-857.
10. Kazienko P., Musiał K., Kajdanowicz T., Multidimensional Social Network and Its Application to the Social Recommender System, IEEE Transactions on Systems, Man, and Cybernetics - Part A: Systems and Humans, 2009, in press.
11. Keles Ay., Kolcak M., Keles Al., The adaptive neuro-fuzzy model for forecasting the domestic debt, Knowledge-Based Systems, 21(8), 2008, pp. 951-957.
12. Kuncheva L., Combining Pattern Classifiers. Methods and Algorithms, John Wiley & Sons, Inc., 2004.
13. Lin P.-C., Chen J.-S., A genetic-based hybrid approach to corporate failure prediction, International Journal of Electronic Finance, 2(2), 2008, 241-255.
14. Pelleg D., Moore A. W., X-means: Extending K-means with Efficient Estimation of the Number of Clusters, International Conference on Machine Learning, Morgan Kaufmann Publishers Inc., San Francisco, USA, 2000, pp. 727-734.
15. Ravi V., Kurniawan H., Nwee P., Kumar R., Soft computing system for bank performance prediction, Applied Soft Computing, 8(1), 2008, pp. 305-315.
16. Rud O., Data Mining Cookbook. Modeling Data for Marketing, Risk, and Customer Relationship Management, John Wiley & Sons, Inc., 2001.
17. Swanson N. R., White H., A model selection approach to assessing the information in the term structure using linear model and the artificial neural network, Journal of Business and Economics Statistics, 13, 1995, pp. 265–275.
18. Zurada J., Lonial S., Comparison Of The Performance Of Several Data Mining Methods For Bad Debt Recovery In The Healthcare Industry, The Journal of Applied Business Research, 21(2), 2005, pp. 37-53.