Ensembles of Locally Linear Models: Application to Bankruptcy Prediction

Laura Kainulainen, Qi Yu, Yoan Miche, Emil Eirola, Eric Séverin and Amaury Lendasse

***Abstract*— The bankruptcies of companies have been predicted with numerous methods. In this paper, the ensemble of Locally Linear model is compared to Linear Discriminant Analysis, Least Squares Support Vector Machines and Optimally Pruned Extreme Learning Machines. To create the ensemble, diffrerent basis for the locally linear models as well as different combinations of variables are used in order to obtain enough diversity between the models. The obtained models are combined into the final model by solving a least-squares non-negative constraints problem. The model is tested on a Polish bankruptcy data set and the results discussed also from the point of view of importance of the variables.**

I. Introduction

Bankruptcies are not only financial but also individual crises which affect many lives. Although unpredictable things may happen, bankruptcies can be predicted to some extent. This is important for both the banks and the investors that analyze the companies, and for the companies themselves.

The aim of this paper is to see, whether the ensembles of Locally Linear models combined with forward selection of the variables perform better than three comparison methods: Linear Discriminant Analysis, Least Squares Support Vector Machines and Optimally Pruned Extreme Learning Machine. They form a good basis for comparison, since LDA is a widely spread technique in the financial tradition of bankruptcy prediction, LSSVM is an example of Support Vector Machine classifiers and OP-ELM is actually a neural network. Since all the possible combinations of the variables cannot be evaluated due to time constraints, forward selection may offer a fast and accurate solution for finding suitable variables, especially combined with ensembles. According to Polikar [1], the main idea in ensemble modeling is to combine several classifiers in order to make one better classifier. The underlying assumption is that single classifiers make errors on different instances.

The ensembles are formed by creating different classifiers of Locally Linear models based on the K Nearest Neighbor method. Diversity of the classifiers is obtained with different values for K and using different variables selected with forward selection. The classifiers are merged together by solving a non-negative least-square constraints problem. A Polish bankruptcy data set is used to test the models. It consists of 120 companies [2].

Laura Kainulainen, Qi Yu, Yoan Miche, Emil Eirola and Amaury Lendasse are with the Aalto University School of Science and Technology, Faculty of Information and Natural Sciences, Konemiehentie 2, 02150 Espoo, Finland.

Eric Séverin is with the University of Lille 1,Dept GEA,Bâtiment SHS n°3, BP 179, 59653 Villeneuve d'Ascq cedex,France.

In section 2, the methods used are explained. In the following part, the experiments and the results obtained are presented. Finally, the importance of the results and further work is discussed.

II. The methodology

This section presents first the reference methods, Linear Discriminant Analysis, Least Squares Support Vector Machines and Optimally Pruned Extreme Learning Machine. Second, the Locally Linear models and the methods of merging them into ensembles are discussed. Also the problem of variable selection is covered. Finally, the methods that were used to estimate the performance of the developed models are presented.

A. Linear Discriminant Analysis

In Linear Discriminant Analysis, the main idea is to calculate a score that would describe the risk of a company to go bankrupt, and classify the scores to bankrupted and healthy according to the chosen threshold. This score is calculated as a linear combination of the explanatory variables. That is to say, each variable is given a weight and then summed. The weights are defined to separate the means of the two classes [3]. The whole idea with discriminant analysis is to give more weight to the variables that separate best the means of the two groups and are the most similar within the groups. Altman also tested, whether the year, when the data has been collected, has influence on the prediction performance. He concluded that even though the accuracy is lower, a dataset collected two years prior to the bankruptcy can be used for the prediction [4].

B. Least Squares Support Vector Machines

Support Vector Machines is a widely spread technique that aims to find a hyperplane that maximizes the margin between two classes. In non linear cases a kernel, i. e. a mapping from the original input space into a high dimensional space is used in order to obtain a problem that could be solved again. Thanks to the dual transformation of the problem, the calculation can be simplified. A good presentation of Support Vector Machines can be found in [5]. Least Squares Support Vector Machines develop the method further by replacing the quadratic programming problem by a set of linear equations. [6], [7]

C. Optimally-Pruned Extreme Learning Machine

Optimally Pruned Extreme Learning Machine is based on the Extreme Learning Machine algorithm. ELM builds

a single hidden layer feedforward neural network (SLFN) with random hidden nodes. OP-ELM ranks the nodes based on Multiresponse Sparse Regression algorithm and prunes them using the results of leave-one-out validation. [8], [9]

D. Locally Linear models

The idea with Locally Linear models is that linear regression is performed for each sample of the data set, based on its K Nearest Neighbors [10]. The KNN algorithm is based on the idea that the K nearest neighbors of a certain sample are used for defining the class of that sample. The sample is labeled to the class which dominates among these neighbors. In this case the distance between two samples was defined as an Euclidean distance although there are other possibilities [11]. However, here the Locally Linear regression is used to predict the class of each sample. The nearest neighbors are used only as a basis to build the regression model. The value of K, meaning how many neighbors are used is at least the number of dimensions d plus one, because otherwise linear regression could not be performed [12]. In these experiments, the maximum number of K is $d + 50$ due to computational time constraints. The dimension d, meaning the number of variables used, changes according to the phase of the forward selection.

To estimate how well each model would perform when shown to completely new data, leave-one-out cross-validation is used. K-fold cross-validation is a technique where the dataset is divided into *K* blocks, and each of the blocks is of size *N/K*, if N is the total number of samples. Each of the blocks is used in turn as a calibration set and the rest *K-1* blocks as a training set. The leave-one-out method is a special case of K-fold cross-validation, where the training set consists of all samples except one, which is used for calibration. It means that the *K* is equal to *N* [13]. In this case, the leave-one-out cross-validation contributes to building a more accurate ensemble, since the models that are estimated to perform the best with new data are favored in the ensemble formation. This also reduces the risk of overfitting.

E. The ensembles of several classifiers

The main idea with the ensembles of several classifiers is that several classifiers are created and then combined into one model. As a result, the process to create the ensembles of classifiers consists of two key components: the diversity of individual classifiers and method of combining the classifiers obtained. It is important to create enough diversity between individual classifiers so that they make errors on different instances. In other words, the decision boundaries of classifiers should be different. This diversity can be obtained in several ways [1].

The following sections present three ways to create diversity between classifiers. First option is to use different values of K in the K Nearest Neighbor. Second, the model is built on different variables. Third, different K in the K Nearest Neighbor and different variables are used in a row. The section II-E.4 presents the second aspect of creating ensembles: merging individual classifiers.

1) Different values of K in the K Nearest Neighbor create diversity: One way to create diversity is to use different values of K to create different classifiers. This means that in Locally Linear models the neighborhood on which the linear regression is built varies. For example, for one data point, we used $d + 1$, $d + 2$, $d + 3$ etc. neighbors. Figure 1 illustrates this idea. The different values of K, $d+1$, $d+2 \ldots d+limit$, are used to create different models M_1, $M_2 \ldots M_n$.

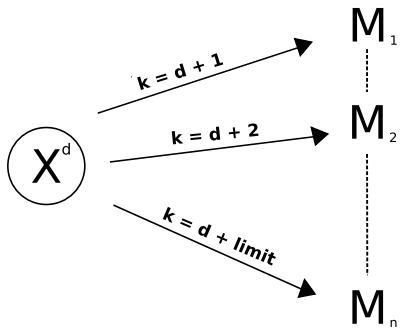

Fig. 1: Creating diversity with different values of K.

2) Different combinations of variables create diversity: Another way to create diversity is to use different combinations of variables. In this case, fixed K is used. Nevertheless, the models differ because they use different combinations of variables. Figure 2 explains this principle. Different variable sets $1, 2, \ldots, n$ are used to create different models M_1, $M_2 \ldots M_n$.

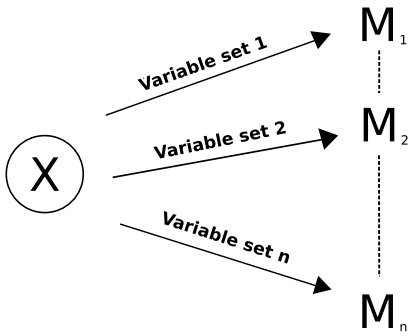

Fig. 2: Creating diversity with different combinations of variables.

How are these variable sets chosen? Forward search is used in this paper, because it enables to keep the number of the variables rather small, which improves interpretation possibilities. Naturally, there are also other possibilities, such

as random subspaces [1]. In forward search, the models are first built on all possible variables, meaning that each model uses one variable. The best variable is chosen. Second, rest of the variables are combined with the variable that was chosen from the first round. The best combination is saved. On every round, one more variable is added to the combination [14].

Figure 3 represents the forward method. One of the five variables is chosen and all the variables in time are added to the set and tested. The set that obtains the highest accuracy (percentage of correct classification) is chosen for a basis of the next round.

3) Different values of K and different combinations of variables create diversity: The two previous sections presented how to use different values of K in the K Nearest Neighbor and different combinations of variables obtained with forward search in order to create the diversity of the classifiers that are merged to ensembles. However, it is possible to use both methods in series. First, models that are based on different variables are created. With forward search, a set of variables is chosen, on which we build models that use different K:s. The K varies between $d+1$ and $d+limit$, d being the dimensions and $limit$ being 50 in this case. These models, which vary in the K used, are merged into ensemble. Many sets of variables are used as a basis and the procedure is repeated several times, which means that in the end there are several ensembles. Second, the ensembles that were obtained in the first phase are combined. This means that in the first phase the variety of the classifiers comes from the different values of K used. In the second phase, the models, which are actually ensembled from the first phase, vary in the variables used. This implies that we obtain the variety of models from two sources: K in the K Nearest Neighbor and the variables used.

Figure 4 illustrates this situation. Different variable sets are used as a basis for Locally Linear ensembles. In an ensemble of Locally Linear models, diversity is created with different values of K, meaning that the number of neighbors used changes. When they are combined, we obtain the estimates $\hat{y}_{LOO}$. These models are combined again into the final model, $\hat{y}$.

4) Combining different models into ensembles: The second aspect of creating the ensembles of classifiers is the method of merging several classifiers. There are several ways to classify these methods, for example to partition them to classifier fusion and selection techniques, or to trainable and non-trainable practices. The method used also depends on the type of the output. If the output consists of class labels, methods such as majority voting or weighted majority voting, are useful. For continuous outputs, many kinds of algebraic combiners, such as weighted average, or decision templates can be used [11], [1]. It must be noted, though, that the continuous outputs can be converted to labeled output simply by using an appropriate threshold. Nevertheless, the choice of the threshold is not obvious.

In this study, to create ensembles, the nonnegative least-squares constraint problem between different classifiers is solved by using the Non-Negative constrained Least-Squares (NNLS) algorithm [15]. It can be considered as an algebraic combiner. According to Miche et al. [12], the advantage of this method is that it is efficient and fast. The leave-one-out outputs of each method, as seen in section II-D, combined with the positivity constraint also reduce the risk of overfitting. The method defines a positive coefficient for each model. The coefficients are defined so that the accuracy of the ensemble would be maximized. It must be emphasized that the coefficients are not defined based on the training data, but the leave-one-outputs, which make the model less prone to overfitting. This idea is also visualized in the Figure 4, where the $\hat{y}_{LOO}$ represent the leave-one-out outputs that are combined into the final model $\hat{y}$.

F. Estimating the performance of the ensemble

The main idea in estimating the performance of the method is to divide the data set into training, validation and testing sets. The models are built in the training phase based on the information that the training set contains. The results are validated and the best model chosen. Finally, the model is tested in a test set that was not used for building the model. However, bankruptcy prediction data sets are often rather small because they are laborious, and above all, expensive to obtain. This makes the performance estimation challenging. As a result, Monte-Carlo cross-test is used for the testing. The leave-one-out cross-validation is used with the Locally Linear models, because then the merging into ensembles is more accurate: the models that are estimated to perform best on the new data are favored.

Monte-Carlo methods refer to various methods. In this study, Monte-Carlo cross-test is adopted. It consists of two steps. First, the data set is divided into training and testing sets. The training set is formed by drawing without replacement a certain number of samples. The testing set comprises the rest of the samples. Second, the model is trained with the training set and then tested with the testing set. These steps are repeated several times [16]. In this case, the training set contains 75% of the samples and the testing set the rest. These two steps are repeated 750 times due to time limitations.

III. Experiments

A. Data set

The data set used in this paper was developed by Wiesław Pietruszkiewicz. It contains 240 cases of which 112 are bankrupted companies and 128 healthy. In total there are 120 companies, because the data comes from two years in a row. The possible bankruptcy occurred from two up to five years after the observations [2]. The 30 variables consist of ratios of different financial variables. They are presented in Table I. A comparison between several datasets would have been favorable, but unfortunately bankruptcy datasets are expensive to obtain, they are not online, and the credit data sets available were not suitable for further analysis of the variables that was made.

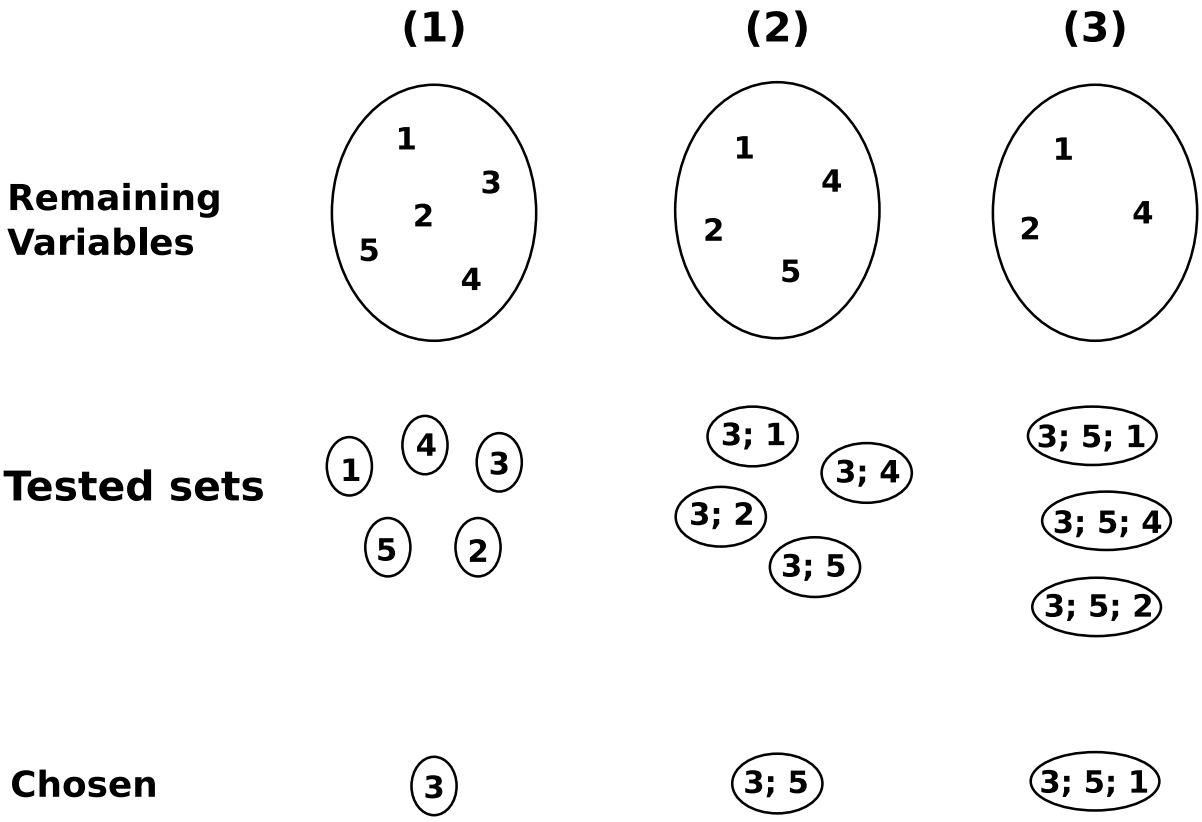

Fig. 3: Forward selection: The sets of variables chosen from the previous round are used as a basis of the following round.

The data is divided into three categories. The first represents how the profit is formed and allocated (5 and 6), the second highlights the financing (for instance 26 and 27), and the third represents the profitability (for instance 2, 13 and 14).

B. Results

The Locally Linear ensemble models are compared to the reference methods. From the Figure 5, we can notice that already with two variables selected with the forward selection, more accurate results are obtained than with Linear Discriminant Analysis. It must be noted that the horizontal line describing the accuracy of Linear Discriminant Analysis is obtained with all the variables. The percentage of correct classification for OP-ELM was 71.71 and for LSSVM 73.45. Thus Locally Linear ensembles obtain better accuracy with 2 and 4 variables, respectively. With Locally Linear ensemble, the optimum of correct classification is obtained with seven variables. However, the results with 6, 8 and 9 variables are very similar.

What are the variables that are chosen the most often? Variables 24, 9 and 13 – all these indicators are variables representing the economic profitability – are chosen the most often as the first variable in forward search, in 49%, 27% and 18% of the cases respectively. They are most often combined with variable 18. This variable stands for the turnover of assets, that is to say how assets are being used to produce revenues. Figures 6, 7 and 8 represent these pairs of variables. They describe the repartition of the companies into bankrupted and healthy ones if the classification is performed only based on these variable pairs. From these figures it can be seen that the variable pairs make sense in a way that the classes are rather separable with them.

The seven variables that are most often obtained are, in descending order of importance, 18, 24, 5, 6, 9, 20 and 27. The frequencies of these variables to be chosen amongst the 9 first variables in forward search are 71%, 52%, 50%, 49%, 43%, 40% and 36 %.

This set of variables can be divided into three main groups: economic profitability (variables 9, 18, 24), financial structure (27), the use of assets (that is to say their ability to generate sales) (variables 18 and 20), and business cycle (variables 5 and 6). These two last variables represent the needs of financing for business cycle. We can note that the working capital is independent of the methods of value fixed assets and depreciation and amortization. However, working capital can be influenced by inventory valuation methods.

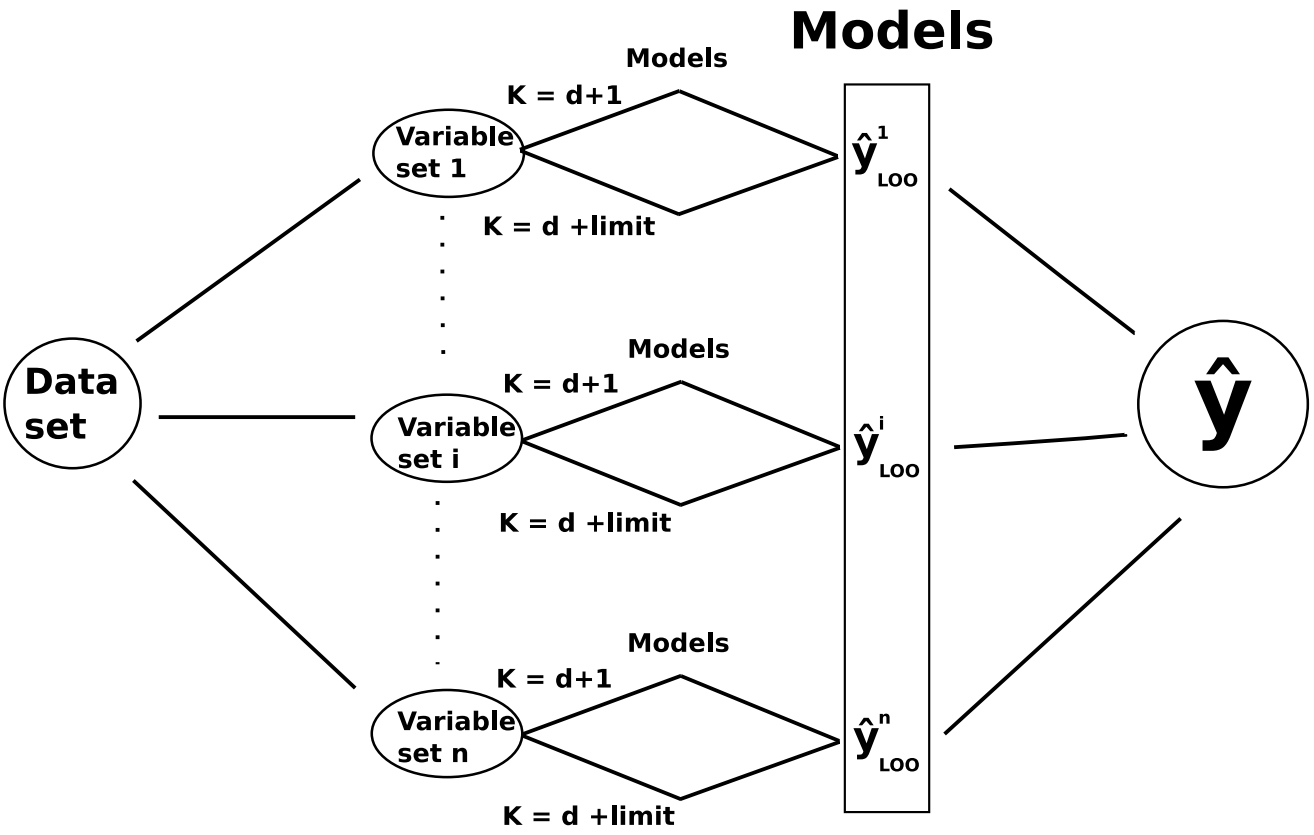

Fig. 4: Two ensembles in a series with diversity from two sources: K Nearest Neighbors and variable sets.

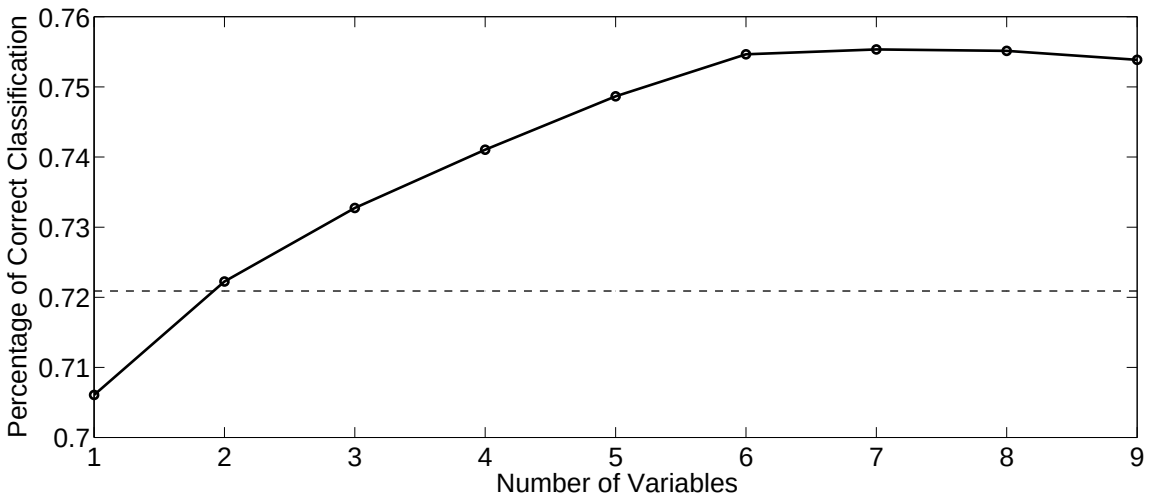

Fig. 5: Correct Classification of Ensemble of Locally Linear models compared to Linear Discriminant Analysis.

Most interestingly, if the results obtained in this study are compared with Altman's results, we find the same groups of variables even if the retained indicators are not the same. In Altman's Linear Discriminant Analysis, we find that the groups of variables are the same. The difference comes from the variables. For economic profitability, the variable used by Altman is 29, for financial structure, the variable is equity divided by total debt, an indicator very similar with

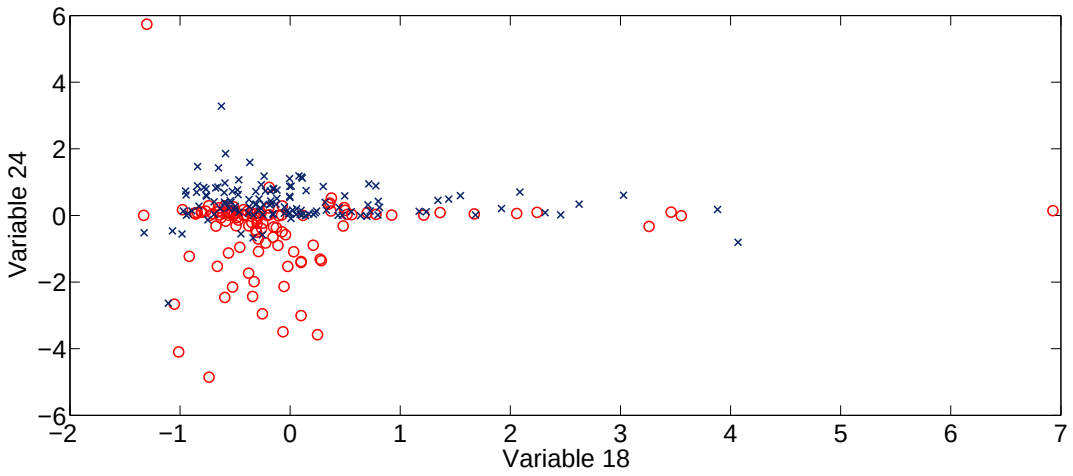

Fig. 6: Classification Based on Variables 18 and 24.

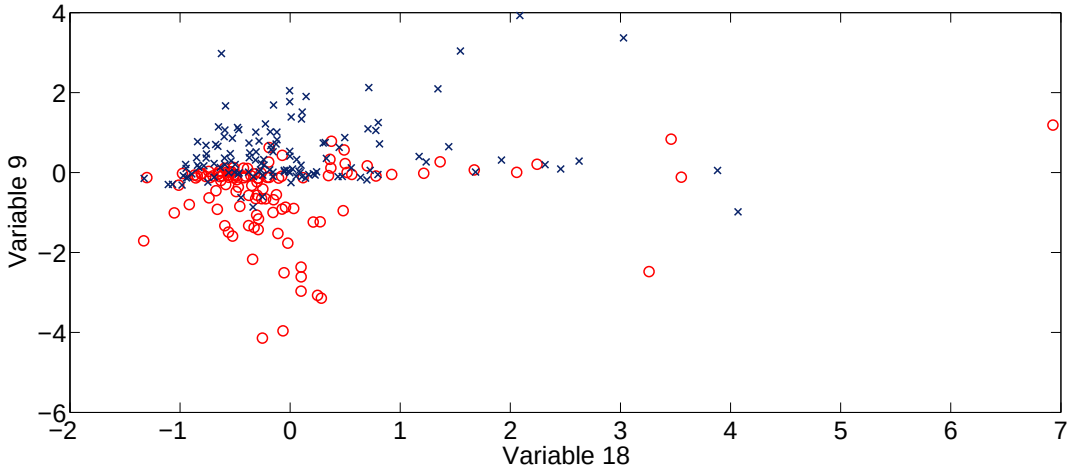

Fig. 7: Classification Based on Variables 18 and 9.

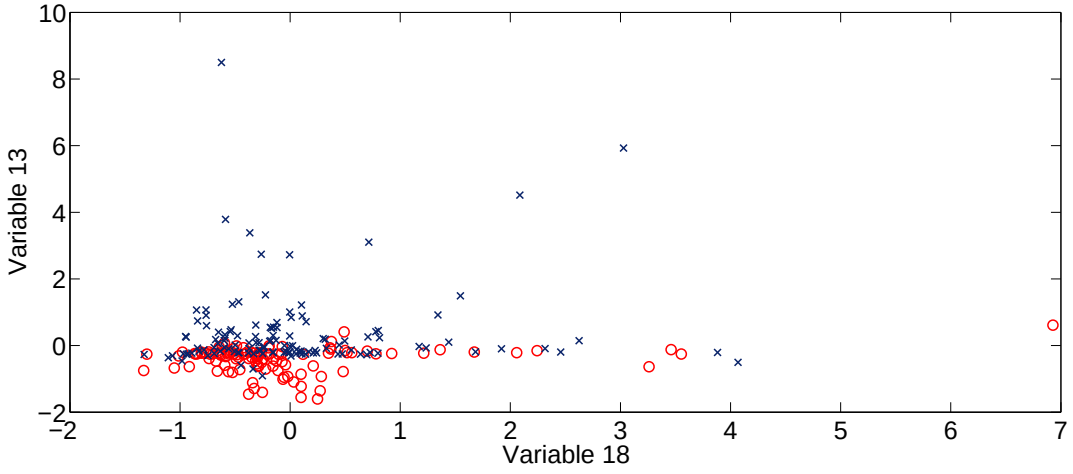

Fig. 8: Classification Based on Variables 18 and 13.

variable 26 in the database used. Finally, for business cycle the variable retained by Altman is the variable 17. In other

TABLE I: The variables used in the experiments

Number	Variable
X1	cash/current liabilities
X2	cash/total assets
X3	current assets/current liabilities
X4	current assets/total assets
X5	working capital/total assets
X6	working capital/sales
X7	sales/inventory
X8	sales/receivables
X9	net profit/total assets
X10	net profit/current assets
X11	net profit/sales
X12	gross profit/sales
X13	net profit/liabilities
X14	net profit/equity
X15	net profit/(equity + long term liabilities)
X16	sales/receivables
X17	sales/total assets
X18	sales/current assets
X19	(365*receivables)/sales
X20	sales/total assets
X21	liabilities/total income
X22	current liabilities/total income
X23	receivables/liabilities
X24	net profit/sales
X25	liabilities/total assets
X26	liabilities/equity
X27	long term liabilities/equity
X28	current liabilities/equity
X29	EBIT (Earnings Before Interests and Taxes)/total assets
X30	current assets/sales

words, the results obtained are in line with Altman but even if the groups are the same, the measurements of these variables are not. That is why choosing variables to design a model is not a secondary task.

IV. Discussion

Based on the performed tests, we can conclude that the ensembles of Locally Linear models combined with forward search can perform better than Linear Discriminant Analysis and OP-ELM with only two variables and better than LS-SVM with four variables. The classification accuracy increases at least until 7 variables and 75.5 % of correct classification. The method is also reasonably fast: the selection of 7 variables and the model building takes about 20 minutes from one core of an I7-920 2.66 GHz processor. An LS-SVM method without variable selection takes 2 minutes, but combined with similar variable selection, the training of a classifier would take approximately 7 hours.

The seven variables that are chosen the most often are 18, 24, 5, 6, 9, 20 and 27. These variables can be divided into three groups: economic profitability, financial structure and business cycle. Interestingly, these groups are the same that Altman used in his research. However, the variables used are not the same.

The advantage of the ensembles of Locally Linear models is that they are accurate yet fast to build. Also the method seems to be able to choose variables that are considered as important in the previous research, which is important for the interpretation of the results. However, further research on other datasets is needed.

Acknowledgment

The authors would like to thank Dr. Wiesław Pietruszkiewicz for the datasets.

References

[1] R. Polikar, "Ensemble based systems in decision making," *IEEE Circuits and Systems Magazine*, vol. 6, no. 3, pp. 21–45, 2006.

[2] W. Pietruszkiewicz, "Dynamical systems and nonlinear kalman filtering applied in classification," in *Proceedings of 2008 7th IEEE International Conference on Cybernetic Intelligent Systems*. IEEE, 2008, pp. 263–268. [Online]. Available: http://www.pietruszkiewicz.com

[3] R. A. Fisher, "The use of multiple measurements in taxonomic problems," *Annals of Eugenics*, vol. 7, pp. 179–188, 1936.

[4] E. I. Altman, "Financial ratios, discriminant analysis and the prediction of corporate bankruptcy," *Journal of Finance*, vol. 23, no. 4, pp. 589–609, September 1968.

[5] N. Cristianini and J. Shawe-Taylor, *An Introduction to Support Vector Machines and Other Kernel-Based Learning Methods*. Cambridge University Press, Cambrige U.K., 2000.

[6] J. A. K. Suykens and J. Vandewalle, "Least squares support vector machine classifiers," *Neural Processing Letters*, vol. 9, pp. 293–300.

[7] J. A. K. Suykens, T. Van Gestel, J. De Brabanter, B. De Moor, and J. Vandewalle, *Least Squares Support Vector Machines*. World Scientific, 2002.

[8] G.-B. Huang, Q.-Y. Zhu, and S. Chee-Kheong, "Extreme learning machine: Theory and applications," *Neurocomputing*, vol. 70, pp. 489–501, December 2006.

[9] Y. Miche, A. Sorjamaa, and A. Lendasse, "OP-ELM: Theory, experiments and a toolbox," in *LNCS - Artificial Neural Networks - ICANN 2008 - Part I*, ser. Lecture Notes in Computer Science, R. N. Vera Kurková and J. Koutník, Eds., vol. 5163/2008. Springer Berlin / Heidelberg, September 2008, pp. 145–154.

[10] G. Bontempi, M. Birattari, and H. Bersini, "Local learning for data analysis," in *Proceedings of the 8th Belgian-Dutch Conference on Machine Learning, Benelearn'98*, F. Verdenius and W. van den Broek, Eds. Wageningen, The Netherlands: ATO-DLO, 1998, pp. 62–68.

[11] L. I. Kuncheva, *Combining Pattern Classifiers, Methods and Algorithms*. John Wiley & Sons Inc., Hoboken, New Jersey, 2004.

[12] Y. Miche, E. Eirola, P. Bas, A. Lendasse, and M. Verleysen, "Ensemble modeling with a constrained linear system of leave-one-out outputs," 2010, unpublished, admitted to ESANN 2010, European Symposium on Artificial Neural Networks.

[13] R. Polikar, "Bootstrap-inspired techniques in computation intelligence," *Signal Processing Magazine, IEEE*, vol. 24, no. 4, pp. 59–72, July 2007.

[14] F. Rossi, A. Lendasse, D. François, V. Wertz, and M. Verleysen, "Mutual information for the selection of relevant variables in spectrometric nonlinear modelling," *Chemometrics and Intelligent Laboratory Systems*, vol. 80, pp. 215–226, 2006.

[15] C. L. Lawson and R. J. Hanson, *Solving leas squares problems*, ser. Classics in Applied Mathematics 15. Society for Industrial and Applied Mathematics, 1995.

[16] A. Lendasse, V. Wertz, and M. Verleysen, "Model selection with cross-validations and bootstraps - application to time series prediction with RBFN models," in *ICANN 2003, Joint International Conference on Artificial Neural Networks, Istanbul (Turkey)*, ser. Lecture Notes in Computer Science, O. Kaynak, E. Alpaydin, E. Oja, and L. Xu, Eds., vol. 2714. Springer-Verlag, June 26-29 2003, pp. 573–580, iSSN: 0302-9743 (Print) 1611-3349 (Online).