Forecasting the bankruptcy of companies: the study on the usefulness of the random subspaces and random forests methods

Barbara Pawełek[1], Dorota Grochowina[2]

Abstract

The subject matter of the paper falls into the mainstream of research on an important economic issue, which is the bankruptcy of companies. The issues related to the bankruptcy of companies have been widely discussed in the economic literature. Works are being continued on the methodology of forecasting the bankruptcy of companies, including the use of Data Mining methods for this purpose. The purpose of the paper is to present the results of empirical studies on the usefulness of the random subspaces and random forests methods for forecasting the bankruptcy of companies in Poland. Both of the aforementioned methods belong to the multi-model approach to data analysis. The random subspaces method relies on the random selection of variables, while the random forests methods uses both the random selection of variables, as well as the random selection of observations. The analysis was performed for balanced sets of companies. The usefulness of the random subspaces and random forests methods for forecasting the bankruptcy of companies was assessed based on the values of the classification accuracy measures for companies from the test part of the input set of objects. The basis for building the set of objects was a set of companies active in the industrial processing sector in Poland in the years 2013-2014. The financial data were taken from the Emerging Markets Information Service website. The computations were done in the R program.

Keywords: *bankruptcy, forecasting, random subspaces method, random forests method, classification accuracy*
JEL Classification: C380, C520, G330

1 Introduction

Establishing new enterprises and closure of the business by some of the existing companies are typical phenomenons for the market economy. Among enterprises, both finishing their activity in the market and continuing their business, entities can highlighted that are subject to court's declaration of bankruptcy by liquidation or bankruptcy with a possibility of concluding an agreement (restructuring proceedings). These companies are colloquially known as the bankrupts. The phenomenon of bankruptcy of enterprises is an important economic issue. Practitioners and researchers are interested in forecasting the bankruptcy risk of companies due to social and economical issues that emerge as a result of such a phenomenon. Issues related to forecasting the bankruptcy of enterprises have been widely

[1] Corresponding author: Cracow University of Economics, Department of Statistics, 27 Rakowicka Street, 31-510 Kraków, Poland, e-mail: barbara.pawelek@uek.krakow.pl.
[2] Warsaw School of Economics, Collegium of Economic Analysis, 6/7 Madalińskiego Street, 02-513 Warszawa, Poland, e-mail: Dorota.Grochowina@doktorant.sgh.waw.pl.

discussed in the economic literature (e.g. Pociecha et al., 2014). Works on the methodology of forecasting the bankruptcy of companies are being continued, including the use of the Data Mining methods (e.g. Min, 2016a; Min, 2016b; Panov and Džeroski, 2007; Pawełek and Grochowina, 2016; Virág and Nyitrai, 2014).

The aim of this paper is to present the results of empirical studies on the usefulness of the random subspaces and random forests methods for forecasting the bankruptcy of enterprises in Poland. The added value of the paper consists in the presentation of results obtained using the selected Data Mining methods for forecasting the bankruptcy of companies operating in the industrial processing sector in Poland, along with the verification of the hypothesis that the values of the classification accuracy measures for the particular considered methods, calculated on the basis of the testing part, come from the populations that have the same average values. The computations were done in the R program, first of all using the packages 'randomForest' (Liaw and Wiener, 2002), 'rpart' (Therneau et al., 2015), 'PMCMR' (Pohlert, 2014).

2 Data and the research procedure

The analysis used a set containing 7223 companies operating in the industrial processing sector in Poland. Among the considered objects were 42 companies that were subject to court's declaration of bankruptcy in 2014 or 2015 (bankrupts, B). The financial data related to the years 2013 and 2014, i.e. they were obtained from the financial statements which were published a year before the bankruptcy. The database also contained 7181 "healthy" companies, i.e. those continuing business activity in 2014 and 2015 (non-bankrupts, NB). The selection of "healthy" companies for the database was based, inter alia, on their core business activity and the availability of financial data for 2013 and 2014. The data were taken from the Emerging Markets Information Service website (https://www.emis.com/pl). The companies' bankruptcy risk was forecast with one year in advance.

The input dataset was used to generate balanced research sets that contained 84 companies (42 B – ½ of total and 42 NB – ½ of total)[3]. The "healthy" companies were included in the research sets through the random selection, repeated 30 times. Each of the thirty research sets was randomly divided (100 times) into a training part (⅔ of total) and a testing part (⅓ of total)[4]. In the analysis, the dependent variable had the zero-one characteristic and received

[3] The considerations presented in this paper are a pilot project. Similar considerations will be made for the unbalanced sets of objects.

[4] The arbitrary proportions and the number of repetitions were adopted for data set division into the training part and testing part.

category "1" in the case of companies that declared bankruptcy in 2014 or 2015, and category "0" for "healthy" companies. 16 financial ratios were used as the independent variables, divided into the following groups: liquidity ratios (3 variables), liability ratios (4 variables), profitability ratios (5 variables), productivity ratios (4 variables).

The methods of random subspaces (Ho, 1998) and random forests in the Forest-RI version (Breiman, 2001), based on the classification tree CART (Breiman et al., 1984)[5] were used in the analysis. Both of the aforementioned methods belong to the ensemble approach to data analysis. The ensemble approach consists in aggregating M base models D_1, ..., D_M into one aggregated model D^*, in order to improve the forecasting accuracy (Gatnar, 2008). The random subspaces method relies on the random selection of variables that are used to the base models, while the random forests method uses both the random selection of the variables and the random selection of observations. The analysis was carried out for one hundred base models (M = 100). The classification of companies into two groups: entities without bankruptcy risk in a year (group "NB") and entities with bankruptcy risk in a year (group "B")[6] was carried out based on each of one hundred base models created for each of three thousand training sets. The aggregation of one hundred base models, i.e. the combination of the forecasting results based on the base models, was performed according to the majority voting method, and resulted in 3000 aggregated models (Gatnar, 2008).

The evaluation of the usefulness of the random subspaces and random forests methods for forecasting the bankruptcy of companies in Poland was performed on the basis of the values of the following classification accuracy measures calculated for companies included in the testing set:

- total error (percentage of companies incorrectly classified into "B" or "NB" groups);
- error type I (percentage of the bankrupts classified into "NB" group);
- error type II (percentage of "healthy" companies classified into "B" group).

The classification accuracy of the random subspaces and random forests methods was compared with the classification accuracy of the single classification tree CART, which is one of the popular methods for forecasting the bankruptcy. 3000 single classification trees were created, one for each of the analysed training sets. The input set of the independent variables included all the considered financial indicators.

[5] Default settings of the given package were left during calculations using `rpart` and `randomForest` functions. Changing some parameters may help to achieve better results than those presented in this paper.

[6] 30 research sets divided 100 times into the training part and the testing part.

Using univariate ANOVA, it was planned to verify the hypothesis that the values of the classification accuracy measures for the particular methods, calculated on the basis of the testing sets, came from populations that had the same average values. The Shapiro-Wilk (Shapiro and Wilk, 1965), Levene and Brown-Forsythe (Brown and Forsythe, 1974) tests were used in order to check if the ANOVA assumptions regarding the normality of the variable distribution in the sets and the equality of the variable variance in the sets, were met. Due the fact that the ANOVA assumptions had not been met, it was decided to use a non-parametrical variance analysis. The Kruskal-Wallis test (Kruskal and Wallis, 1952), followed by the post-hoc Conover test (Conover and Iman, 1981) and post-hoc Dunn test (Dunn, 1964), were used, whereas, the Bonferroni multiple testing correction was considered in the post-hoc tests.

A common feature of the random subspace and random forests methods is the reduction of the variables space dimension. The calculations were performed for all possible variants of the reduction of the set of variables, i.e. the number of variables ranging from 1 to 15. Researchers involved in the ensemble approach to the data analysis often point to the usefulness of a formula derived from the information theory for the determination of the dimension of the reduced space (e.g. Amit and Geman, 1997; Breiman, 2001). According to this approach, the number of variables should be equal to the integer number from the interval $[\log_2 K, \log_2 K + 1)$, where K is the number of input variables (Cover and Thomas, 2006). In the research carried out by the authors, $K = 16$, therefore the results for four variables are presented.

3 Results of empirical research

In the first phase of the research, on the basis of each of the thirty research sets and each of one hundred training sets, the bankruptcy of the companies was projected with one year in advance, using the single classification tree, random subspace and random forests methods. Afterwards, the classification errors were calculated (total, type I and type II) for the companies included in the respective testing sets. It resulted in 27 000 values of errors (30 research sets × 100 divisions into training and testing parts × 3 Data Mining methods × 3 error types). The obtained results provided the basis for analysis performed in the subsequent phases.

The second phase of the analysis consisted in checking, whether the values of the classification accuracy measures for the considered methods, calculated on the basis of the test sample, came from the populations that had the same average values. For this purpose, it was checked whether the ANOVA assumptions regarding the normality of the variable distribution in the sets and the equality of the variable variance in the sets, had been met.

The Shapiro-Wilk test was used for each of the thirty research sets and each of the three considered methods, based on the error values (respectively: total, type I and type II), which occurred during the classification of objects belonging to the one hundred test samples analysed. Table 1 (columns 2-4) contains numbers indicating how many times *p*-value in the Shapiro-Wilk test was lower than 0.05 for 30 research sets. The results show that the assumption of the normality of distribution of the total, type I and type II errors, which define the classification effectiveness of the considered methods (as measured based on the test sample) in most of the analysed cases at a significance level of 0.05, has not been met.

Group	Shapiro-Wilk test			Levene test	Brown-Forsythe test	Kruskal-Wallis test
	Tree	RS	RF			
Total	20	25	30	11	11	28
Bankrupts	29	29	30	20	20	28
Non-bankrupts	24	30	30	26	26	27

Symbols: Tree – single classification tree, RS – random subspace method, RF – random forests method.

Table 1. The number of research sets (max = 30) for which the *p*-value was lower than 0.05 respectively in the Shapiro-Wilk (columns 2-4), Levene (column 5), Brown-Forsythe (column 6), Kruskal-Wallis (column 7) tests.

Then the assumption about the equality of the variables variances in the groups was verified. The Brown-Forsythe and Levene tests were used for this purpose for each of the thirty research sets. Table 1 contains the numbers indicating how many times *p*-value in the Levene (column 5) and Brown-Forsythe (column 6) tests was lower than 0.05 for 30 research sets. The test results at a significance level of 0.05 show that the assumption of the equality of the variance of the type I and type II errors that define the classification accuracy of the considered methods (as measured based on the test set) have not been met in most of the analysed cases. In the group including bankrupts and non-bankrupts, the test results at a significance level of 0.05, in most cases show that there are no grounds to reject the null hypothesis that assumes the equality of the variance of type I and type II errors that define the classification effectiveness of the considered methods (as measured based on the test set).

In the third phase of the analysis, as the ANOVA assumptions had not been met by the considered variables in the majority of the analysed cases, it was decided to use a non-parametrical variance analysis. The Kruskal-Wallis test was employed. Table 1 contains

numbers indicating how many times *p*-value in the Kruskal-Wallis test (column 7) was lower than 0.05 for 30 research sets. In most cases, at a significance level of 0.05, the null hypothesis was rejected in favour of an alternative hypothesis that stated that for at least two out of three considered methods, the values of the classification accuracy measures, as calculated based on the test set, come from populations that had different average values.

In the fourth phase of the analysis, the post-hoc Conover and Dunn tests with the Bonferroni correction for the multiple testing were used to determine which sets of the results did not come from the populations that had the same average values. Table 2 contains figures indicating how many times *p*-value in the post-hoc Conover (columns 2-4) and post-hoc Dunn (columns 5-7) tests was lower than 0.05 for 30 research sets.

Group	**post-hoc Conover test**			**post-hoc Dunn test**		
	Tree – RS	**Tree – RF**	**RF – RS**	**Tree – RS**	**Tree – RF**	**RF – RS**
Total	26	24	21	26	24	18
Bankrupts	22	22	14	22	22	14
Non-bankrupts	22	19	15	22	18	13

Symbols: Tree – single classification tree, RS – random subspace method, RF – random forests method.

Table 2. The number of research sets (max = 30) for which the *p*-value was lower than 0.05 respectively in the post-hoc Conover (columns 2-4), post-hoc Dunn (columns 5-7) tests.

The results in table 2 show that in most cases, at a significance level of 0.05, it should be considered that the values of the classification accuracy measures, as calculated based on the test set, for the single tree and the random subspace method and for the single tree and the random forests method come from the populations having different average values. The same conclusion can be made for the results for the random subspace method and the random forests method in the group of bankrupts and non-bankrupts. In the case of separate sets of bankrupts and non-bankrupts, the results are not conclusive. At the significance level of 0.05, the number of cases in which it was necessary to reject the null hypothesis and the number of cases in which there were no grounds to reject the null hypothesis were similar.

In the last, fifth phase of the analysis it was verified, which of the considered methods was characterized by higher classification accuracy, as measured based on the test set. The arithmetic average the errors (respectively: total, type I and type II) in the classification of the objects included in the one hundred test sets analysed, was calculated for each of the thirty

research sets and each of the three considered methods. Then, the methods were compared in terms of the obtained result. Table 3 (columns 2-4) contains numbers indicating how many times the difference between the average value of the errors occurred for the thirty analysed research sets.

Group	Arithmetic average			Ranks average (post-hoc Conover test)		
	Tree > RS	**Tree > RF**	**RF > RS**	**Tree > RS**	**Tree > RF**	**RF > RS**
Total	30	30	17	30 (26)	29 (24)	17 (14)
Bankrupts	29	29	18	29 (22)	28 (22)	17 (9)
Non-bankrupts	28	26	18	28 (22)	25 (19)	18 (7)

Symbols: Tree – single classification tree, RS – random subspace method, RF – random forests method.

Table 3. The number of the research sets (max = 30) for which the specific difference between the arithmetic average values occurs (columns 2-4) and the number of the sets for which the specific difference between the average rank values occurs, and additionally for which the *p*-value in the post-hoc Conover test was lower than 0.05 (values in brackets) (columns 5-7).

The analysis of the results presented in table 3 (columns 2 and 3) leads to conclude that the use of the random subspace and random forests methods in forecasting the bankruptcy of enterprises in the industrial processing sector in Poland, in majority of cases resulted in an increase of the classification accuracy of the test set objects, compared to the single classification tree. Comparison of the average error values calculated for the random subspace and random forests methods (table 3, column 4), shows the advantage of the random subspace method. However, considering fact that 30 pairs of values were compared, the results should be approached with some reservation.

The Kruskal-Wallis, post-hoc Conover and post-hoc Dunn tests are based on the ranks and their arithmetic averages. The calculating of the average value of the ranks means an operation unacceptable for the measurements on the ordinal scale and involves a strengthening of the measurement scale. However, in order to determine which of the considered methods was characterized by higher classification accuracy, as measured based on the test set, in combination with examination of significance of the differences between the pairs of the average values of the populations, it was decided to calculate the average rank

values and to compare the methods in terms of obtained results. Table 3 (columns 5-7) contains numbers indicating how many times the difference between the average rank values occurred for the thirty research sets analysed. The numbers in parentheses indicate how many times the specific differences between the average rank values are accompanied by the *p*-value of the post-hoc Conover test smaller than 0.05. The discussions considered only the post-hoc Conover test due to the similarity between the results of this test and the results of the post-hoc Dunn test.

The obtained results strengthen the conclusion concerning the improvement of the classification accuracy of the objects from the test set as a result of using the random subspace and random forests methods, compared to the single classification tree. In addition, the believe about the need for exercising caution when attempting to indicate a more effective method between the random subspace and random forests methods has been confirmed.

Conclusions

On the basis of the results of the research carried out it can be found that the use of the random subspaces and random forests methods in forecasting the bankruptcy of companies helps to improve the classification accuracy of test set objects compared to the single classification tree.

The results do not provide the grounds for distinguishing any of the analysed methods (i.e. either the random subspaces method or the random forests method), as the one the application of which more strongly favours the improvement in the classification accuracy of the test set objects. This conclusion concerns in particular a situation where sets of bankrupts and non-bankrupts are considered separately. Therefore, no justification has been found for using the more complex method, which is the random forests method, compared to the random subspaces method,. This problem requires further analysis, including e.g. a change of the parameters in the `rpart` and `randomForest` functions from the default ones to those dedicated to the phenomenon under study.

The above conclusions have been formulated based on the results of analyses carried out on the basis of actual data. However, it should be borne in mind that every empirical study has specific limitations. The presented study was focused on companies active in the industrial processing sector in Poland. The financial data related to the years 2013 and 2014, and the forecast was performed one year in advance. The research set was balanced. The training part included ⅔ of the companies of the research set and the test part included the rest of the objects, i.e. ⅓ of all companies.

In further studies, the authors intend to repeat the analysis for unbalanced sets. It is also planned to include the Forest-RC type of the random forests method and the *v*-times cross validation in the analysis.

Acknowledgements

Publication was financed from the funds granted to the Faculty of Management at Cracow University of Economics, within the framework of the subsidy for the maintenance of research potential.

References

Amit, Y., & Geman, D. (1997). Shape quantization and recognition with randomized trees. *Neural Computation, 9*, 1545–1588.

Breiman, L. (2001). Random Forests. *Machine Learning, 45*, 5–32.

Breiman, L., Friedman, J., Stone, C. J., & Olshen, R. A. (1984). *Classification and regression trees*. CRC Press, Boca Raton.

Brown, M. B., & Forsythe, A. B. (1974). Robust Tests for the Equality of Variances. *Journal of the American Statistical Association, 69(346)*, 364–367.

Conover, W. J., & Iman, R. L. (1981). Rank Transformations as a Bridge Between Parametric and Nonparametric Statistics. *The American Statistician, 35(3)*, 124–129.

Cover, T. M., & Thomas, J. A. (2006). *Elements of Information Theory*. Second Edition, A Wiley-Interscience publication, A John Wiley & Sons, Inc., Hoboken, New Jersey.

Dunn, O. J. (1964). Multiple Comparisons Using Rank Sums. *Technometrics, 6(3)*, 241–252.

Gatnar, E. (2008). *Podejście wielomodelowe w zagadnieniach dyskryminacji i regresji [The multi-model approach to discrimination and regression problems]*. PWN Scientific Publishers, Warsaw.

Ho, T. K. (1998). The random subspace method for constructing decision forests. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 20(8)*, 832- 844.

Kruskal, W. H., & Wallis, W. A. (1952). Use of Ranks in One-Criterion Variance Analysis. *Journal of the American Statistical Association, 47(260)*, 583–621.

Liaw, A., & Wiener, M. (2002). Classification and Regression by randomForest. *R News, 2(3)*, 18–22.

Min, S.-H. (2016a). A genetic algorithm-based heterogeneous random subspace ensemble model for bankruptcy prediction. *International Journal of Applied Engineering Research, 11(4)*, 2927–2931.

Min, S.-H. (2016b). Integrating instance selection and bagging ensemble using a genetic algorithm. *International Journal of Applied Engineering Research, 11(7),* 5060–5066.

Panov, P., & Džeroski, S. (2007). Combining Bagging and Random Subspaces to Create Better Ensembles. In: Berthold, M. R., Shawe-Taylor, J., & Lavrač, N. (eds.), *IDA 2007, LNCS 4723,* Springer-Verlag, Berlin Heidelberg, 118–129.

Pawełek, B., & Grochowina, D. (2016). The random subspace method in the prediction of the risk of bankruptcy of companies in Poland. In: Malina, A., Węgrzyn, R. (eds.), *Knowledge – Economy – Society. Challenges and Development of Modern Finance and Information Technology in Changing Market Conditions,* Foundation of the Cracow University of Economics, Cracow, 25–34.

Pociecha, J. (ed.), Pawełek, B., Baryła, M., & Augustyn, S. (2014). *Statystyczne metody prognozowania bankructwa w zmieniającej się koniunkturze gospodarczej [Statistical methods of forecasting the bankruptcy in changing economic conditions*]. Foundation of the Cracow University of Economics, Cracow.

Pohlert, T. (2014). The Pairwise Multiple Comparison of Mean Ranks Package (PMCMR). *R package,* URL: http://CRAN.R-project.org/package=PMCMR.

Shapiro, S. S., & Wilk, M. B. (1965). An Analysis of Variance Test for Normality (Complete Samples). *Biometrika, 52(3/4),* 591–611.

Therneau, T., Atkinson, B., & Ripley, B. (2015). rpart: Recursive Partitioning and Regression Trees. *R package version 4.1-10,* URL: http://CRAN.R-project.org/package=rpart.

Virág, M., & Nyitrai, T. (2014). The application of ensemble methods in forecasting bankruptcy. *Financial and Economic Review, 13(4),* 178–193.