ISSN: 0974-5823 Vol. 7 No. 5 May, 2022

International Journal of Mechanical Engineering

Feature Selection Using the Mahalanobis Distance for Bankruptcy Prediction

Maria Elena BRUNI [1]

[1] Department of Mechanical, Energy and Management Engineering, University of Calabria, Rende (Cosenza), Italy.

Patrizia BERALDI [2]

[2] Department of Mechanical, Energy and Management Engineering, University of Calabria, Rende (Cosenza), Italy.

Gianpaolo IAZZOLINO [3]

[3] Department of Mechanical, Energy and Management Engineering, University of Calabria, Rende (Cosenza), Italy.

Dinara ZHAISANOVA [4]

[4] Department of Management , The Higher School of Economics and Business,

Al-Farabi Kazakh National University, Almaty, Kazakhstan.

Abstract - In recent years credit risk analysis, credit scoring and bankruptcy prediction have become hot topics in the credit industry. The issues are addressed by employing techniques that intensively use a huge amount of high-dimensional data to make financial predictions. This richness of data brings serious challenges, since irrelevant and redundant information may greatly degrade the performance of learning algorithms. Many studies consider feature selection as a pre-processing step necessary to increase efficiency in learning tasks, improve learning performance like predictive accuracy, and reduce the computational burden of learning algorithms. This paper moves along this direction, and applies a feature selection method based on the use of the Mahalanobis distance on bankruptcy prediction. Computational experiments are carried out on European large-sized firms categorized into the data base as ”Other services” using the Amadeus/Orbis database as data source.

Keywords : *Feature selection, Mahalanobis, Financial distress, Financial ratios, Amadeus*/Orbis

JEL Classification : C10, G32, G33, M41

INTRODUCTION

Financial institutions face every day the problem of determining the creditworthiness of business companies they lend to. In particular, financial actors strive to make appropriate lending decisions ([3]) by analyzing business plans and financial information, to select borrowers with a clear ability to repay.

The availability of a huge amount of data has stimulated the use of various statistical and machine learning techniques to make financial predictions concerning the likelihood that a firm may step into bankruptcy, or to quantitatively determine the insolvency risk. These studies routinely analyse financial measures of the chosen dataset to disclose their importance and the potential impact on bankruptcy or financial distress.

As there is no general agreement on the best set of financial ratios for bankruptcy prediction and credit scoring, a non-statistical approach is often used to analyse the collected input variables (or features) and to eventually discard those elements which are loosely correlated with a distressed or insolvency status. This process is known in the literature as the feature selection process, enables reductions of high-dimensional datasets.

Among the others, one of the most used methods for feature selection is the wrapped-based approach. It wraps the feature selection algorithm around a classification algorithm, usually a Support Vector Machine (SVM), which is used as a black box in the whole process. This approach is especially useful when a fitness function cannot be easily expressed with an exact mathematical expression.

Different feature selection methods have been proposed in the literature for predicting bankruptcy or to address the credit scoring problem. In particular, in [5] the author attempted to analyze the prediction accuracy of models designed with various classification techniques and variable selection methods. In [17] bankruptcy prediction is made by means of five statistical based feature selection methods, which are analyzed and compared. Zhou et al. [21] proposed a new approach for feature selection based on direct search and features ranking technology to construct hybrid SVM models for bankruptcy prediction.

In this research, we propose the use of a particular distance metric (the Mahalanobis distance) as fitness function within a genetic algorithm based wrapper feature selection method for bankruptcy prediction.

LITERATURE REVIEW

Feature selection methods have emerged as crucial tools in applications where the curse of dimensionality of data affects the classification algorithm's performance. The process explores the space of feature subsets with the aim of scoring the subsets and choosing the best one in terms of classification. The approaches proposed in the literature are mainly three: the filters, the wrappers and the embedded methods ([4]).

Filter approaches ([16]) are independent of any machine learning algorithm and analysze the property of the data, regardless the chosen classifier. In the wrapper approach ([9]), the selection of a set of features is considered as a search problem, and different combinations of features are explored and evaluatedon the basis of the classifier performance. Exhaustive search on the space of feature subset can be computationally very intensive for large datasets. For this reason the exploration id performed heuristically, through search algorithms, as sequential search, which start with an empty set (full set) and add features (remove features) until the maximum objective function is obtained, and evolutionary algorithms [11]. Amongst the latters, genetic algorithms (GAs, for short) are very popular. They have shown their efficacy in various areas as computer vision/image processing, text mining, bioinformatics [1,2, 7]. Examples of the application of the GA to feature selection for bankruptcy prediction can be found in Min and Lee [10]. Recently, Zelenkov et al. [20] combined a wrapper method based on GA in a two-step method; at the first stage, the significant features for ordinary classifiers are selected which are combined into a voting ensemble in the second stage whose weights are determined again by using the GA.

In the GA heuristic, the individuals belonging to a given generation are modified by the appplication of genetic operators, which mutate and recombine the genetic information of the population to produce a new generation, hopefully better than the previous one. The best individuals, selected on the basis of a so-called fitness function, survive for the next generation, whilst non promising individuals are discarded.

When GA is applied in bankruptcy prediction, as fitness function can be used the classification accuracy, for instance obtained by applying a SVM method [18]. A straightforward application of a GA on feature selection might not produce the expected result, since the algorithm outputs a high number of features. Rejer [14] addressed this issue proposing a variant of the GA, called GA with Aggressive Mutation (GAAM), where individuals contain a fixed number of features set by the user. In order to identify the feature subset of minimum cardinality, the algorithm should be run with a decreasing number of genes several times, leading to a very computational intensive procedure. A variant of the GAAM, referred to as GA with melting individuals (GAMI, for short) was proposed in [15] to overcome this problem. At the core of the GAMI, the GAAM is called for a predefined number of iterations. The number of features is minimized during the execution of the GA, by randomly removing from each individual one feature in successive iterations of the algorithm (see Figure 1).

As the GAAM, the GAMI also uses as fitness function the classifier accuracy, tested with tenfold cross-validation. Even though this method has been proven to provide accurate results, it has a main drawback - the slow performance-, since each subset of features should be evaluated by the classifier. The aim of this paper is to propose a new method for feature selection that uses the Mahalanobis distance as the fitness function within a GA-based wrapper approach to forecast bankruptcy.

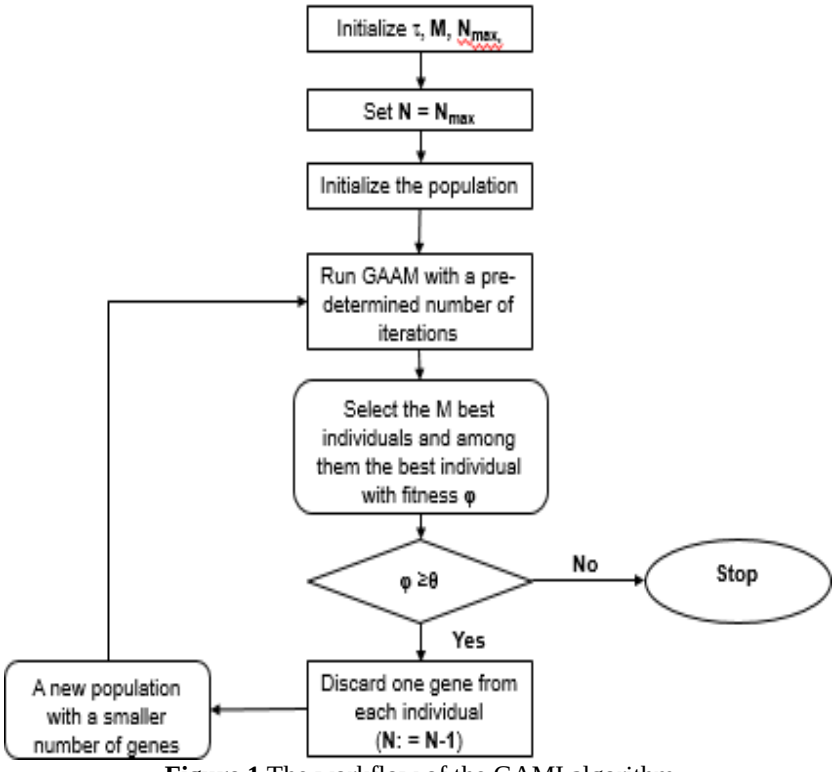

Figure 1 The workflow of the GAMI algorithm

METHOD

Our method is based on a tailored variant of the GAMI, where the Mahlanobis distance is used as fitness function. The resulting variant will be referred in the following as GAMD (acronym for Genetic Algorithm with Mahlanobis Distance). . We note that the use of distance metrics in machine learning algorithms is not a novelty, see for instance [19].

As any GA, the genetic information of each individual should be properly encoded to guarantee the successful application of the metaheuristic. Other relevant aspects are a careful design of the genetivc operators and the right choice of the fitness function. In our case the fitness function, (that will be indicated in the following with φ^{MAL}) is evaluated on the basis of the Mahalanobis distance function. The Mahalanobis distance is a well-known statistical distance function among classes of data points. In particular, if we have two groups of data with mean $x\hat{}^i$ and $x\hat{}^j$, respectively, and a covariance matrix Σ, the Mahalanobis distance is given by

$$\varphi^{Mal} = \sqrt{(x\hat{}^i - x\hat{}^j)^T \Sigma^{-1} (x\hat{}^i - x\hat{}^j)}.$$

It reduces to the Euclidean distance metric when the features are uncorrelated and hence deemed equally important and independent from others. A previous research ([12]) showed that the Mahalanobis distance is good at identifying the importance of correlated features and the distance between groups of data. In particular, it was assessed that the performance of the SVM notably deteriorates when the Mahalanobis distance decreases. This suggested the use of this distance metric for feature selection: the idea was to evaluate each subset of features by means of the distance between positive and negative data in the dataset. To encode the genetic information of each individual, we used a chromosome with N genes (where N is set at the beginning of the algorithm), each representing a given feature. The initial number of features is P and therefore, each individual corresponds to a subset of N_{MAX} features.

Figure 2 reports the workflow of the GAMD. The algorithm starts with a random population of M individuals coding different subsets of features. In particular, for each individual, values from the set 0, 1, 2, . . . P , are randomly generated. We assume that a value of 0 means that no feature has been selected, hence reducing the total number of features.

Then, the GAAM is applied for a given number of iterations IT_{max}. In each iteration the individuals are crossed-over and mutated. The classic one-point crossover is applied on the individuals from the current population population P^{curr}, creating M new individuals, stored in P^{cross}. An aggressive mutation is also performed on the current population, by randomly creating NM new individuals, each differing from the father for one gene.

The newly created population (with $M + NM + M$ individuals) (containing the individuals from the mother population P^{curr}, as well as from the mutated population P^{mut} and from the new generation obtained by applying one-point crossover P^{cross}) is ordered on the basis of the fitness function, with the aim of discarding the worst $NM + M$ individuals (Step 17).

Algorithm 1: The workflow of the GAMD

1 **Input:** The maximum number of iterations IT_{max}, the maximum number of genes of an individual N_{max} and the threshold τ.
2 **Initialization:** Set the number of genes $N := N_{max}$, $best, currentbest := null$; $P^{curr}, P^{cross}, P^{mut}, P^{best} = \emptyset$.
3 Create an initial population of M individuals randomly choosing N values from the set $\{0, 1, 2, \ldots, P\}$ and store them in P^{curr}
4 **repeat**
5 $it = 0$
6 $P^{it} = P^{curr}$
7 **while** $it \leq IT_{max}$ **do**
8 $P^{cross}, P^{mut} = \emptyset$
9 **for** $i \in P^{it}$ **do**
10 Apply 1 -point crossover operator to i, obtaining a new individual and store it in P^{cross}
11 **end**
12 **for** $i \in P^{it}$ **do**
13 **for** $g = 1, \ldots, N$ **do**
14 Assign a random value from the interval $\{0, 1, 2, \ldots, P\}$ to the gene g and save the individual i as a new individual in P^{mut}
15 **end**
16 **end**
17 Select the best M individuals from $P^{mut} \cup P^{it} \cup P^{cross}$ according to the evaluated fitness function ϕ^{MAL} and store them in P^{curr}
18 it:=it+1
19 **end**
20 $currentbest = argmin_{i \in P^{curr}} \phi_i^{MAL}$
21 **if** $\phi^{MAL}_{currentbest} > \phi^{MAL}_{best}$ **then**
22 best:=currentbest
23 **end**
24 **for** $i \in P^{curr}$ **do**
25 Randomly select one gene to discard
26 **end**
27 Set the number of genes $N := N - 1$
28 **until** $\phi^{MAL}_{currentbest} < \tau$;
29 Return $best$

Figure 2 The GAMD algorithm

Then, the fitness value of the best individual is selected and evaluated against a threshold τ. We define τ as the Mahalanobis distance between two different groups of data coming from samples classified by the classifier as positive and negative in the training data.

If $\varphi^{MAL}_{currentbest} \geq \tau$ The number of features is iteratively decremented in each individual, where a gene is randomly selected and discarded , otherwise, the algorithm terminates.

EXPERIMENTAL RESULTS

In order to assess the efficiency of the proposed approach, we extract information on a set of 553 European large-sized firms from the Amadeus/Orbis database, which contains comprehensive information on millions of companies across the world. Our primary sample includes firms categorized into the database as "Other services", that are businesses which do not fit into Orbis categories and, significantly, include all high technology-based businesses. Companies with no recent financial data and Public authorities/States/Governments were then excluded. A total of 123 bankruptcy firms was identified. A firm has been considered bankrupted also if it was dissolved (bankruptcy or liquidation), inactive (no precision), active (default of payment) or in liquidation.

We have considered a comprehensive list of 45 relevant accounting-based and market-based variables listed in Table 1. In line with the current BASEL III practice, we have evaluated for each model the out-of-sample prediction performance.

A personal computer with a CPU Intel(R) Core(TM) i5-2450M @2.50Ghz ws used for the computational experiments. The GAMD parameter setting is as follows: $IT_{max} = 100$, M=10, Nmax = 6, and $\tau = 2.34$.

Six features were selected by the GAMD (NDM, EBITDA%, ROA, ROI, SR, IATA) with an accuracy classification accuracy, evaluated through SVM of 91,5%. ROA and ROI are typical indicators of profitability for the firm. It is very reasonable that it is included in the predictors of financial distress. The solvency (SR) is important for evaluating financial default. The EBITDA is the most important indicator of the operating activity of the firm. There are also non-conventional variables that result as important predictors of default. By non-conventional variables we mean variables that are not traditionally used for financial statements analysis and for prediction of financial distress. The following non-conventional variables were selected: NDM (Number of Directors and Managers) learly connected to human resource and IATA (Intangible Assets / Total Assets) related to investments towards intangible assets.

Table 1. Variables

Variable	Description	Variable	Description
OR	Operating revenue (Turnover)	TA	Total assets
SF	Shareholders funds	NE	Number of employees
FA	Fixed assets	IA	Intangible fixed assets
CA	Current assets	LTD	Long term debt
NCL	Non-current liabilities	L	Loans
CL	Current liabilities	S	Sales
EBIT	Earnings before Interest and Taxes	NI	Net income
CE	Cost of employees	DA	Depreciation & Amortization
IP	Interest paid	CF	Cash flow
EBITDA	Earnings Before Interest,Taxes, Depreciation and Amortization	R&D	Research & Development expenses
EPS	Earnings per share	NDM	Number of directors & managers
NP	Number of patents	NT	Number of trademarks
EBITDA%	EBITDA/Sales	EBIT%	EBIT/Sales
NI%	Net income/Sales	R&D%	Research & Development expenses/Sales
ROE	Net income/Shareholders funds	D	Debt (Long term debt + Loans)
PC	Permanent capital	IC	Invested capital (Shareholders funds+ Debt)
ROA	EBIT/ Total assets	ROI	EBIT/ Invested capital
EBITDATA	EBITDA/ Total assets	AT	Asset turnover (Sales/ Total assets)
ROD	Interest paid/ Debt	DI	Debt/ Invested capital
SR	Solvency ratio (Permanent capital/ Fixed assets)	CR	Current ratio (Current assets/ Current liabilities)
VA	Value added (EBITDA+ Costs of employees)	ACE	Average cost for employee
VACE	Value added/ Costs of employees	IATA	Intangible assets / Total assets
E	Experience		

We have compared the results of the GAMD approach with Boruta, a recursive feature selection algorithm and a plain version of the GA. Boruta is a wrapper feature selection algorithm which is based on random forest variable importance measure. The importance of each independent variable (feature) is evaluated by fitting a random forest on an extended dataset created with the addition of shuffled copies of all features (shadow features). In every iteration, if a real feature has a better Z-score than the best of the shadow features is kept, otherwise it is removed since it is deemed unimportant.

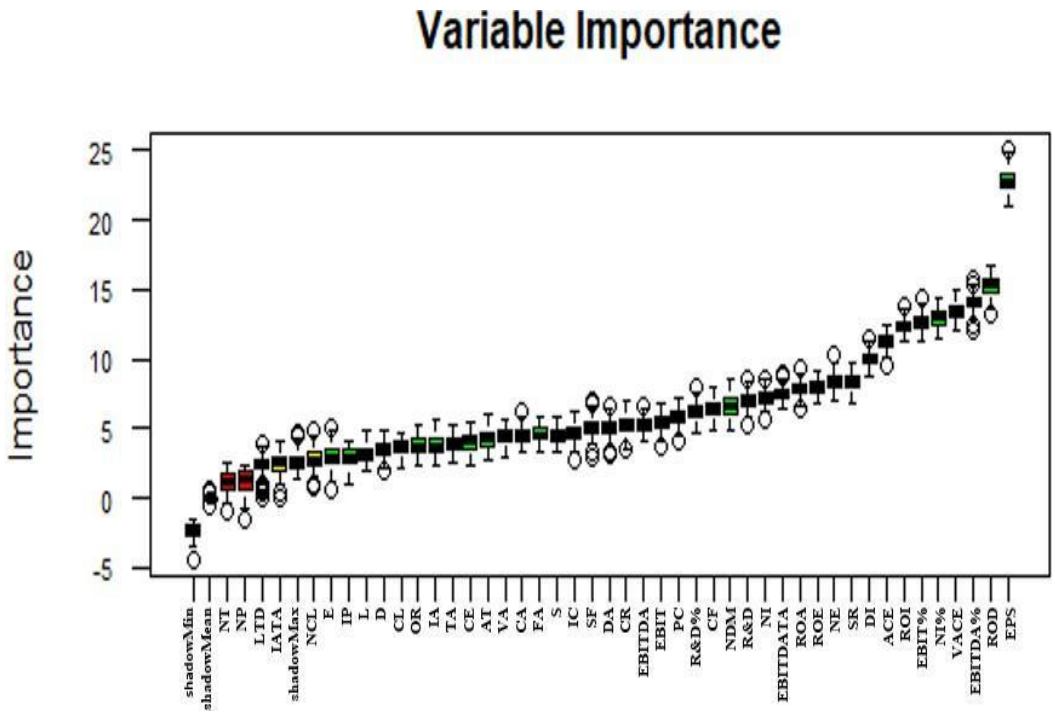

Figure. 3 Boruta result plot

In our study, variable selection is performed by using the R package Boruta.

In Figure 3 the variable importance level found by Boruta algorithm is reported, showing colored boxplots of all the attributes plus minimum, average and max shadow score. A green boxplot shows the important features. Tentative attributes are highlighted in yellow, while unimportant features are represented by red boxplots.

According to Figure 3, we can observe that the EPS values can be considered in the highest level of importance, while eight variables can be grouped in the next level of importance. Besides the usual financial related indicators commonly used in financial analysis, we observe the presence of the VACE (VA/Cost of Employees) and the ACE (Average cost of employees) indicators, which are normally used to evaluate the efficiency of human resources. In a balanced scorecard approach to firm performances ([8]), VACE and ACE are included in the Learning and Growth perspective, that is the typical area related to human capital.

To further investigate the importance of the features able to provide good predictive performance, we have also applied recursive feature selection, using the R package Caret, which recursively ranks the predictors and retains only the top ranked ones. The top five variables obtained are EPS, EBITDA%, ROD, VACE, ROI. The SVM classification accuracy reported by using these five features was below 80%. The ReturnOn Debt (ROD) is a rate of return for the bank and represents an average interest rate paid by the firm during the accounting period. The EPS represents the remuneration for shareholders.

As a second set of experiments, we have compared our GAMD with the GA implemented in the R package with a population of 10 individuals and a maximum number of iteration set to 100. The mutation rate is 0.1 and crossover is 0.8. The GA selected 16 features (FA, CL, IP, EPS, EBIT, ROA, SR, VACE, IA, DA, R&D%, NT, ROE, EBITDATA, CR, IATA), with an accuracy ratio of 87,27%.

The interesting results are that, on average, the GAMD with less features outperforms the baseline GA in terms of prediction accuracy. We should also mention that the computational time of the GAMD is one order of magnitude lower than the time taken by the GA implemented in R.

Also in this case, there are non-conventional variables that result as important predictors of default as VACE (VA/Cost of Employees) that is an indicator normally used to evaluate the efficiency of human resources.

Table 2 Probability of default

Company Name	Status	PD (GA features)	PD (GAMD features)
DVRG N.V.	Dissolved (bankruptcy)	24%	28.6%
PERFECT TECHNOLOGIES SA	Dissolved (bankruptcy)	40%	61.5%
CYRANO SA	Dissolved (bankruptcy)	41%	64.7%
CHAUDRONNERIE TUYAUTERIE MONTAGE	Dissolved (bankruptcy)	23%	30.1%
UNI LAND S.P.A.	Bankruptcy	22%	14.5%
ACTA S.P.A.	Dissolved (bankruptcy)	99%	99.5%
INSTITUTUL NATIONAL DE STICLA SA	Bankruptcy	37.6%	45.7%
INTERACTIF DELTA PRODUCTION	Dissolved (bankruptcy)	82.7%	90.8%
HOLOSFIND	Bankruptcy	74.7%	76.9%
MEDICAL DEVICE WORKS	Bankruptcy	37.8%	61.7%
EDF ENERGIES NOUVELLES	Active	0.02%	1.6%
GENMAB A/S	Active	0.09%	4.1%
EI TOWERS S.P.A.	Active	2.59%	5.1%
GAZTRANSPORT & TECHNIGAZ SA	Active	1.7%	0.4%
ALSTRIA OFFICE REIT-AG	Active	8.4%	14.6%
SEDLMAYR GRUND UND IMMOBILIEN KGAA	Active	0.00%	13.1%
VIGO SYSTEM S.A	Active	2.99%	12.98%
EYEMAXX REAL ESTATE AG	Active	9.28%	0.58%

Further validation

Additional experiments have been performed on a subset of firms, with the aim of assessing the quality of the selected features, when used to evaluate the probability of default (PD, for short), a measure often used by practitioners for internal ratings and bankruptcy prediction. The model assumes a logit link between the a set of independent variables and the default event. Then the firm is considered defaulted if $PD \geq 0.5$.

The results, for a balanced sample of 18 defaulted and not defaulted companies not present in the training set, are reported in Table 2. The PD has been evaluated considering the 16 features selected by the GA and the six features obtained by the GAMD. The correct classification of non-bankrupt firms is 100% in both cases. Although there are only two possible outcomes, it is always easier to predict non-bankrupt firms since they are more numerous in the data sets and in the real world.

The correct classification of bankrupt firms with the GA features data is 30%. The GAMD feature selection method improves the classification of bankrupt firms over logistic regression from 30% to 60%. From these findings, we can draw the conclusion that GAMD significantly outperforms all the other methods in terms of classification accuracy.

Table 3 Pros and cons of the methods compared

Algorithm	Language	Results	Advantages	Disadvantages
GA	R	16 features selected	Implementations available in R and Matlab. Accuracy in SVM 87,27%.	Long computational time.
GAMD	Matlab	6 features selected	Fast computational time. High accuracy ratio. Accuracy in SVM 91,5%.	Needs implementation.
Boruta	R	16 features selected reduced to 5	Implementation available in R. Accuracy in SVM less than 80%.	Long computational time.

CONCLUSIONS

In this paper, we have investigated the problem of features selection for bankruptcy prediction proposing an approach based on the Mahalanobis distance. This approach has been compared with the traditional GA-based feature selection and other methods taken from the literature. From the experimental results, we can observe that the GAMD can consistently keep high performance. Moreover, the model outperforms the other methods in terms of computational time. The overall results confirm the great importance of some kinds of traditional financial performance indicators for predicting financial distress, but at the same time highlight the increasing importance of intangible assets for evaluating firm performances. This aspect deserves further attention, and sheds light on interesting future research.

REFERENCES

[1] Brester, C., Semenkin, E., Sidorov, M., Minker, W.: Self-adaptive multi-objective genetic algorithms for feature selection. In Proceedings of International Conference on Engineering and Applied Sciences Optimization, 1838–1846 (2014)

[2] Bruni, M.E., Nguyen Duy, D., Beraldi, P., Violi, A.: The Mahalanobis Distance for Feature Selection Using Genetic Algorithms: An Application to BCI. New Trends in Emerging Complex Real Life Problems. Springer, 73–81

[3] Bruni, M.E., Beraldi, P. and Iazzolino, G.: Lending decisions under uncertainty: a DEA approach. International Journal of Production Research, 52(3), 766-775 (2014)

[4] Chandrashekar, G., Sahin, F.: A survey on feature selection methods, Computers and Electrical Engineering, 40, 16-28 (2014)

[5] Du Jardin, P.: Predicting bankruptcy using neural networks and other classification methods: The influence of variable selection techniques on model accuracy. Neurocomputing (2014), http://dx.doi.org/10.1016/j.neucom.2009.11.034

[6] Eads, D., Hill, D. Davis, S., Perkins, S., Ma, J., Porter, R., Theiler, J.: Genetic al- gorithms and support vector machines for time series classification. 5th Conference on the Application and Science of Neural Networks, Fuzzy Systems and Evolutionary Computation. 74–85 (2002)

[7] Garcia-Nieto, J., Alba, E., Jourdan, L., Talbi, E.: Sensitivity and specificity based multiobjective approach for feature selection: Application to cancer diagnosis. Information Processing Letters, 109(16), 887–896 (2009)

[8] Kaplan, R.S., Norton, D.P.: The Balanced Scorecard: Translating Strategy into Action, Harvard Business Press, Boston, MA, USA (1996)

[9] Kohavi, R., John, GH.: Wrappers for feature subset selection - Artificial intelligence 97(1-2), 273–324 (1997)

[10] Min, S.H., Lee, J., Han, I.: Hybrid genetic algorithms and support vector machines for bankruptcy prediction. Expert Systems with Applications 31, 652-660 (2006)

[11] Mukhopadhyay, A., Maulik, U., Bandyopadhyay, S., Coello, C. A. C.: A survey of multiobjective evolutionary algorithms for data mining: Part I. IEEE Transactions on Evolutionary Computation, 18(1), 4–19 (2014)

[12] Nguyen Duy, D., Nguyen Hoang, H., Nguyen Xuan, H.: The impact of high dimen- sionality on SVM when classifying ERP data-A solution from LDA. Proceedings of the Sixth International Symposium on Information and Communication Technology. ACM, New York, NY, USA, 32-37. http://dx.doi.org/10.1145/2833258.2833290 (2015)

[13] Provost, F., Fawcett, T.: Robust classification for imprecise environments. Machine Learning, 42(3), 203–231 (2001)

[14] Rejer, I.: Genetic algorithm with aggressive mutation for feature selection in BCI feature space. Pattern Anal. Appl. 18, 485-492 (2015a)

[15] Rejer, I.: Genetic Algorithms for Feature Selection for Brain Computer Interface. International Journal of Pattern Recognition and Artificial Intelligence. World Scientific Publishing Company. 29(5), 1559008-1–1559008-27 (2015b)

[16] Sanchez-Marono, N., Alonso-Betanzos, A., Tombilla-Sanroman, M.: Filter Methods for Feature Selection -A Comparative Study, In: Yin H., Tino P., Corchado E., Byrne W., Yao X. (eds) Intelligent Data Engineering and Automated Learning - IDEAL 2007. Lecture Notes in Computer Science, vol 4881. Springer, Berlin, Heidelberg (2007)

[17] Tsai, C. F.: Feature selection in bankruptcy prediction. Knowledge-Based Systems (2009), http://dx.doi.org/10.1016/j.knosys.2008.08.002

[18] Vapnik, V.: Statistical learning theory. NY, Wiley-Interscience (1998)

[19] Yang, L., Jin, R.: Distance metric learning: a comprehensive survey, Technical Report, Michigan State University (2006)

[20] Zelenkov, Y., Fedorova, E., Chekrizov, D.: Two-step classification method based on genetic algorithm for bankruptcy prediction. Expert Systems with Applications, 88, 393-401 (2017)

[21] Zhou, L., Lai, K.K., Yen, J.: Bankruptcy prediction using SVM models with a new ap- proach to combine features selection and parameter optimisation, International Journal of Systems Science (2014), https://doi.org/10.1080/00207721.2012.720293