Credit Risk Modeling of USA Manufacturing Companies Using Linear SVM and Sliding Window Testing Approach

Paulius Danenas and Gintautas Garsva

Department of Informatics, Kaunas Faculty, Vilnius University, Muitines St. 8, LT- 44280 Kaunas, Lithuania
{paulius.danenas,gintautas.garsva}@khf.vu.lt

Abstract. This paper presents a study on credit risk evaluation modeling using linear Support Vector Machines (SVM) classifiers, combined with feature selection and "sliding window" testing approach. Discriminant analysis based evaluator was applied for dynamic evaluation and formation of bankruptcy classes. The research demonstrates a possibility to develop and apply an intelligent classifier based on original discriminant analysis method evaluation and shows that it might perform bankruptcy identification even better than original model.

Keywords: Support Vector Machines, linear SVM, machine learning, credit risk, evaluation, bankruptcy.

1 Introduction and Related Research

Company classification by their risk can be described as one of the key components of credit risk evaluation model. It plays an important role in decision process of acception/rejection projects of credit application. This problem applies through analysis of various financial and other customer data to conclude the final decision. Sophisticated and effective tools to solve task must be developed. The combination of machine learning and statistical techniques might help to minimize the drawbacks of separate techniques and thus develop models which might prove more accurate than common statistical techniques. This research proposes a technique which is based on popular machine learning technique, namely Support Vector Machines. The proposed method is also tested in "sliding window" approach manner, which means that it can be useful to identify more general trends. Moreover, the combination of this method with discriminant analysis (or other similar techniques) might be useful while trying to improve the performance of these methods by identifying the most relevant financial attributes and developing a new classifier based on that particular technique.

The application of intelligent classification techniques in credit risk domain is dated back to 1968 when Altman et. al. [1] applied discriminant analysis. They obtained 96% and 79% accuracy by using two different samples, however, it is

W. Abramowicz et al. (Eds.): BIS 2012, LNBIP 117, pp. 249–259, 2012.

reliable in its predictive ability only in two years, after that the results fall down significantly. Zmijewski [2] applied probit (simple probit and bivariate) and maximum likelihood principles to a set of 40 bankrupt and 800 non-bankrupt companies and a prediction sample of 41 bankrupt and 800 non-bankrupt companies collected from American and New York Stock Exchanges, resulting in 72% accuracy for complete dataset case. Springate [3] developed his model using step-wise multiple discriminate analysis to select 4 ratios which best describe a failing company. It obtained an accuracy rate of 92.5% using the 40 companies tested by Springate; later 83.3% and 88% accuracy rates were reported after testing it with other samples [4]. Ohlson used logit approach to construct his model [5], and he reported accuracy of 96.12%, 95.55% and 92.84% for prediction within one year, two years and one or two years respectively.

Support Vector Machines is applied for efficient classification obtaining results comparable to Neural Networks and other machine learning techniques. As for credit risk domain, they were successfully applied for company failure prediction [6], financial warning prediction [7], to evaluate financial fate of Dotcoms [8], rating companies [9], to estimate probability of default [10], to study credit rating systems [11], capital risk assessment [12]. Lai and Zhou proposed several SVM based methods for various credit risk related tasks, such as identification of high-risk customers [13] or credit scoring [14]. These authors also developed several Least Squares SVM (LS-SVM) based methods, including their developed Weighted LS-SVM techniques [15] [16] or LS-SVM ensemble models [16][17]. LS-SVM integration into credit risk process was also researched by van Gestel et al. in their works [18][19]; they showed that LS-SVM can provide better performance in both complexity and accuracy. These authors also combined it with Bayesian evidence framework for regularization and kernel parameter selection to predict financial distress of Belgian and Dutch firms with middle market capitalization [20].

A model for forecasting changes which combines discriminant analysis technique together with a supervised neural network applied to increase performance in terms of accuracy has been proposed in [21]. This model was applied to forecast changes in discriminant models although it may be applied to forecast changes in ratings or expert evaluations as well.

SVM has been intensively researched in this field with combination with various soft computing techniques; the advantages and disadvantages of these combinations are described in [22]. Danenas and Garsva [23] tried to combine SVM classification technique with discriminant analysis for credit risk evaluation. Their research showed that LIBLINEAR and SMO algorithms are capable to obtain results similar to Vapnik's SVM classifier results. A comparative research of various SVM classifiers by these authors [24] proved that linear SVM classifiers can be a good alternative for credit risk evaluation model development in terms of both complexity and speed, in case there is no need for nonlinear separation using complex kernel functions.

2 The Method

2.1 Description of Algorithms Used in this Experiment

Support Vector Machines. Support Vector Machines is an efficient and effective solution for pattern recognition problem whereas a following minimization problem has to be solved in order to generate weight vector:

$$\min -\sum_{i=1}^{\ell}\alpha_i + \frac{1}{2}\sum_{i=1}^{\ell}\sum_{j=1}^{\ell} y_i y_j \alpha_i \alpha_j K(\mathbf{x}_i, \mathbf{x}_j)$$

$$\textit{subject to } \sum_{i=1}^{\ell} y_i \alpha_i = 0,\ \forall i: 0 \le \alpha_i \le C$$

where the number of training examples is denoted by *l*, training vectors $X_i \in R, i = 1,..,l$ and a vector $y \in R^l$ such as $y_i \in [-1;1]$. α is avector of l values where each component αi corresponds to a training example (x_i, y_i). If training vectors x_i are not linearly separable, they are mapped into a higher (maybe infinite) dimensional space by the kernel function $K(x_i, x_j) \equiv \phi(x_i)^T \phi(x_j)$ where classifier is generated by minimizing an appropriate convex cost function. This can be done in Support Vector Machines (SVM), Least Squares SVM (LS-SVM) [18][19] and other kernel based learning techniques, such as kernel regression or kernel PCA (which is used for extraction of linearly uncorrelated variables). Then the solution is obtained in the dual space from a finite dimensional convex quadratic programming problem for SVM or a linear Karush–Kuhn–Tucker system in the case of LS-SVM, avoiding explicit knowledge of the high dimensional mapping and using only the related positive (semi) definite kernel function [20].

LIBLINEAR. LIBLINEAR is an open source library and a family of linear SVM classifiers for large-scale linear classification which can be very efficient for training large-scale problems. These classification methods do not use kernel functions for transformation into other space which makes it possible train a much larger set much faster.

Formally these algorithms (except Crammer and Singer algorithm) are defined as follows: given training vectors $x_i \in R^n$, i = 1,..,l in two class, and a vector $y \in R^l$ such that $y_i = \{1,-1\}$, a linear classifier generates a weight vector w as the model using a decision function

$$\operatorname{sgn}(w^T x)$$

One-vs-All (OVA) strategy is used for multiclass classification problems; that is, for K-class problem, K binary classifiers are built separating one class from the rest, and the answer is chosen according to the hyperplane which separates the point with the highest confidence from other data points.

An approach proposed by Crammer and Singer for solving an optimization problem is based on multiclass classification, thus it is defined differently [25]: given training vectors $x_i \in R^n$, i = 1,..,l and a vector $y \in R^l$ such that $y_i \in \{1,..,k\}$ a weight vector is generated using

$$\arg\max_{m=1,..,k} w_m^T x$$

Table 1 gives formulations of these algorithms (algorithms and primary optimization problems that are solved); more information is given in [25].

Table 1. Definitions of the algorithms used in research

Algorithm	Minimization problem
L2-regularized L1-loss SVC	$\min_w \frac{1}{2} w^T w + C\sum_{i=1}^{l} (\max(0,1-y_i w^T x_i))$
L2-regularized L2-loss SVC	$\min_w \frac{1}{2} w^T w + C\sum_{i=1}^{l} (\max(0,1-y_i w^T x_i))^2$
L2-regularized logistic regression	$\min_w \frac{1}{2} w^T w + C\sum_{i=1}^{l} \log(1+e^{-y_i w^T x_i})$
L1-regularized L2-loss SVC	$\min_w \lVert w \rVert_1 + C\sum_{i=1}^{l} (\max(0,1-y_i w^T x_i))^2$ ($\lVert \cdot \rVert_1$ defines 1-norm)
Multi-class SVM by Crammer and Singer	$\min_{w_m,\xi_i} \frac{1}{2} \sum_{m=1}^{k} w_m^T w^m + C\sum_{i=1}^{l} \xi_i$ subject to $w_{y_i}^T x_i - w_m^T x_i \geq e_i^m - \xi_i$, $i = 1,..,l$ $e_i^m = 0$, if $y_i = m$ $e_i^m = 1$, if $y_i \neq m$

These classifiers also include a bias term b, which handled by augmenting the weight vector ***w*** and each instance x_i with an additional dimension. An interesting an useful notice is that all these classifiers have a considerably low number of additional parameters (i.e., only cost parameter *C* and bias parameter *b*) which makes it easier to choose appropriate classifier parameters.

2.2 Research Methodology

This research applies modified method proposed in [23][24], using classifiers defined in previous section.. The modified algorithm is defined as follows:

1. Evaluate each financial entry by using discriminant analysis (or any other expert evaluation method, if possible) and compute bankruptcy classes.
2. Eliminate instances which could not be evaluated in Step 1 because of lack of data or division by zero and thus resulted in empty outputs.
3. Remove attributes from the dataset which have less values than specified threshold (70% was considered in this case).
4. Data imputation is performed by filling missing values with average value of particular attribute.
5. Perform the following steps for each $m \in [1, n-k]$, where n is the total number of periods, k is the number of periods which are used for forecasting:
   a. Apply feature selection procedure in order to select the most relevant attributes and reduce number of dataset dimensions;
   b. Perform classifier parameter selection manually or using heuristic procedures;
   c. Train classifier using data from first m periods.
   d. Apply hold-out testing using data from period p, $p \in [m+1, m+k]; p \in \mathrm{N}$.

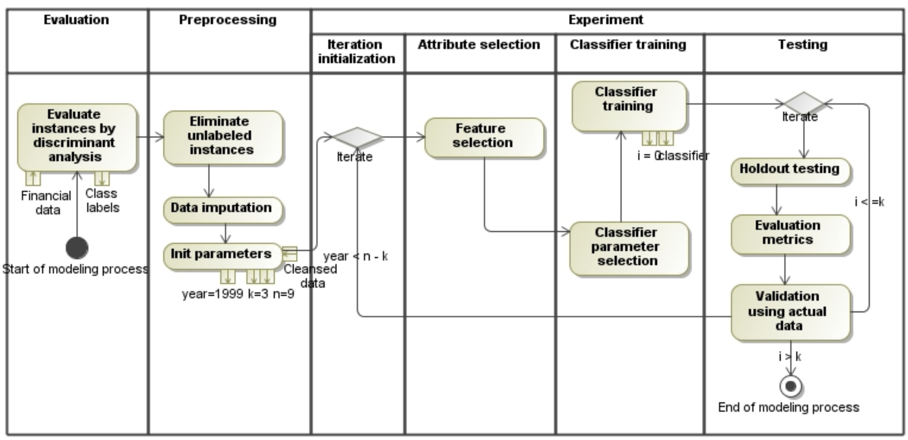

Fig. 1. Workflow of method used in experiment

Figure 1 represents the algorithm graphically as a workflow. The output (for each iteration in experimental stage) is the trained classifier (list of support vectors in case of SVM) and the list of selected attributes.

Finally, to test model performance, an additional step is performed using real bankruptcy data. If applied dataset is in the period $[p_{start}; p_{end}]$, with year p_{end} as the year of last entry in financial history, bankruptcy is known to be occured after the financial history, i.e., on year $p_{end}+1$, $p_{end}+2$,.. $p_{end}+k_y$, with k_y as the maximum number of years during which the company is officially recognized as bankrupt.

Fig. 2. Sliding window approach for testing and bankruptcy identification

Figure 2 gives a graphical overview of overall method, which comprises training, testing and identification (prediction) stages. In this picture, k = 3 and k_y = 3 (this combination is used in experiment). I.e., the data from 1999 to 2005 is used for training, the developed model is used for testing data from year 2006, year 2007 and year 2008 individually). After testing with year 2006, bankruptcy identification is performed on this year: the instance in financial history record representing year p_{end} is labeled as "Risky" (as it was bankrupt), and prediction procedure is performed on the instance.

3 The Experiment

3.1 Research Data

The experiments were made by using data from EDGAR database, manufacturing sector, from year 1999-2008. The initial dataset used in the experiment consists of yearly financial records with 51 financial ratios used in financial analysis; these ratios were computed using original primary financial data from balance and income statement data.

Table 2. Main characteristics of data used in experiment

Year	Entries labeled as		Total entries	No of selected attributes	Bankrupt 1 years after	Bankrupt >1 year after
	Risky (R)	Not risky (NR)				
1999	1312	537	1849	12	-	-
2000	1869	589	2458	15	0	0
2001	1753	672	2425	15	1	0
2002	1709	777	2486	13	3	0
2003	1770	723	2493	14	0	2
2004	1920	637	2557	13	0	1
2005	1964	660	2624	14	3	17
2006	1636	429	2065	14	0	3
2007	1545	393	1938	14	1	13
2008	483	109	592	14	4	1
Total	**15961**	**5527**	**21487**		**12**	**37**

Table 2 presents main characteristics of dataset, including classes formed by evaluation using Zmijewski's score, together with bankruptcy data from UCLA database used to validate the results. UCLA LoPucki database [30] contains bankruptcy data and covers about 50 companies from used dataset. The data from 2000 – 2010 period was applied for validation; instances which represent last entry in financial history were marked as "risky" and were evaluated by developed classifiers.

3.2 Experiment Configuration

Zmijewski's score [2] was used in this research as an evaluator to form class labels and formulate the problem as a classification problem; it was selected because of the origin of the data (which comes from USA and Canada companies). This scoring technique allows to form two groups of companies – companies which are "healthy" (possibly are not going to bankrupt) and "bad' (which might become bankrupt). Zmijewski's score is defined as follows:

Z= -4,336 – 4,513(Net Revenue/Total Assets) + 5,679 * (Total Debt/Total Assets) + 0,004 * (Short Term Assets/ Short Term Assets)*

If Z < 0 then company is considered as "risky"(prone to bankruptcy).

The code and algorithms for the experiments was implemented using Weka framework [28] with LIBLINEAR 1.7. The cost parameter C and bias b were chosen experimentally, by using grid search in range of $C \in [0;100]$ and $b \in [0;1]$. Feature selection was applied for each formed dataset using correlation-based feature subset evaluation [29] to select the most relevant financial ratios.

For the Step 6 of our procedure, it is presumed that bankruptcy might have occurred following the year of the last entry of financial history for particular company, next year or even later (k years after). $k = 3$ is selected in this experiment; thus bankruptcy fact is evaluated here only if it happens during the next 3 years after the last entry in financial records of the company.

The results were evaluated using accuracy, True Positive Rate (TPR) and F-Measure. These metrics are often used in machine learning and more information about them can be found in various sources, such as [27] where these metrics were used to evaluate results.

3.3 Experiment Results

Table 3 presents the classification results - classifier parameters, classification accuracy together with TPR and F-Measure rates for each class. The accuracy is above 90%, which can be considered as very good result. Best results were obtained using different classifiers – Crammer-Singer multiclass SVM showed best performance for 2 analyzed cases, L1 dual linear SVM – for 4 cases and L2 linear SVM, both primal and dual – for last two cases (once per each classifier). Thus different classifiers obtained best results for different periods.

Table 3. Results of experiment

Training period			2000	2001	2002	2003	2004	2005	2006	2007
Structure (parameters)			CS-SVM	L1-LSVM (dual)	L1-LSVM (dual)	CS-SVM	L1-LSVM (dual)	L1-LSVM (dual)	L2-LSVM (primal)	L2-LSVM (dual)
C			20	20	20	15	20	15	15	5
Bias			0.7	1.0	0.7	1.0	0.4	0.7	0.7	1.0
Year 1	**Accuracy**		**96,702**	**96,344**	**95,471**	**95,504**	**91,604**	**93,085**	**92,008**	**92,295**
	TP	**R**	0,973	0,974	0,970	0,965	0,974	0,977	0,971	0,981
		NR	0,951	0,940	0,917	0,925	0,745	0,756	0,724	0,675
	FMeas	**R**	0,977	0,973	0,968	0,970	0,945	0,957	0,951	0,954
		NR	0,941	0,942	0,922	0,911	0,818	0,820	0,789	0,770
Year 2	**Accuracy**		**96,183**	**94,233**	**95,348**	**96,785**	**92,940**	**91,445**	**91,960**	-
	TP	**R**	0,966	0,966	0,972	0,983	0,977	0,966	0,977	-
		NR	0,953	0,938	0,898	0,923	0,749	0,716	0,675	-
	FMeas	**R**	0,972	0,970	0,969	0,979	0,956	0,947	0,952	-
		NR	0,940	0,928	0,906	0,936	0,816	0,775	0,762	-
Year 3	**Accuracy**		**96,032**	**96,286**	**96,710**	**97,389**	**91,291**	**92,127**	-	-
	TP	**R**	0,962	0,970	0,987	0,987	0,964	0,981	-	-
		NR	0,956	0,940	0,908	0,923	0,716	0,667	-	-
	FMeas	**R**	0,972	0,975	0,978	0,984	0,946	0,953	-	-
		NR	0,933	0,927	0,933	0,936	0,772	0,764	-	-

The TPR values for both "risky" (R) and "non-risky" (NR) classes were high (both were over 0.9 in first four periods, and over 0.7 in next periods); this shows that instances for both of these classes were recognized separated and procedures for unbalanced learning were not needed to apply. High F-Measure values which are more suitable for unbalanced learning evaluation also prove this. Parameters C and bias varied; the experiment showed that bias parameter had significant influence thus further research targeted at parameter selection might show even better results.

Table 3 shows that best classification results were obtained while training classifier sequentially with data from first five years (starting with year 1999). Classification resulted in accuracy over 95%.Later it decreased, although the number of instances used for training increased. This might indicate a trend of changes in the data, as well as overall financial situation change; yet the classification performance remained above 90%.

Table 4. Bankruptcy prediction results

Year	Number of actual bankrupt	Original model (Zmijewski)	No of bankruptcies after testing period		
			Year 1	Year 2	Year 3
2002	1	0	0	-	-
2003	3	0	0	0	-
2005	2	0	0	0	0
2006	4	1	1	1	1
2007	1	0	1	0	0
2008	8	6	6	5	5
2009	27	9	18	16	17
2010	3	0	1	1	1
Total:	**49**	**16**	**27**	**23**	**24**

The last step was performed in order to compare the performance of proposed approach with the performance of the original model. Table 4 represents identification (prediction) results. As this table shows, the model developed by the proposed method identified more bankruptcy facts than original Zmijewski model which was used as the evaluator. This might mean that additional statistically selected predictors improved the performance and identified more bankruptcies than the original model in which ratios were selected on the basis of their performance in prior studies. The results varied for each year; yet overall performance was better. Note that Table 2 shows there were far more financial ratios considered relevant by feature selection procedure than the ones that were used in original evaluator. This proves that usage of higher dimensional data and more complex model might result in improved results.

4 Conclusions and Further Research

An approach for credit risk evaluation using linear SVM classifiers, combined with feature selection and sliding window testing is presented in this article. The classifier used here is based on linear SVM classifiers which are perfectly suitable for large scale learning. The developed classifiers were applied for real-world dataset, together with widely applied Zmijewski technique as a basis for output formation. This approach could serve as a alternative tool for company classification in case when there are no actual bankruptcy classes as well as if obtaining them might be a too complicated or expensive. The classifiers were rerun on datasets based on the same principle as described above. Model validation was performed on real bankruptcy list; the obtained results showed that it outperformed the original Zmijewski model.

One of the main problems related to proposed method is possible imbalanced learning arising from the fact that classes are computed dynamically by external evaluator. Although this research did not have to deal with this problem integration of such procedure would be a useful complementary step. This is crucially important in identification of hazardous companies if they are represented by minority entries, as identification of hazardous company might cost more to the creditor than the misidentification of it.

References

1. Altman, E.: Financial ratios, discriminant analysis and the prediction of corporate bankruptcy. The Journal of Finance 23(4), 589–609 (1968)
2. Zmijewski, M.: Methodological Issues Related to the Estimation of Financial Distress Prediction Models. Journal of Accounting Research 22, 59–82 (1984)
3. Springate, G.L.V.: Predicting the Possibility of Failure in a Canadian Firm. Unpublished M.B.A. Research Project, Simon Fraser University (1978)
4. Sands, E.G., Springate, G.L.V., Var, V.: Predicting Business Failures. CGA Magazine, 24–27 (May 1983)
5. Ohlson, J.A.: Financial Ratios and the Probabilistic Prediction of Bankruptcy. Journal of Accounting Research 18(1), 109 (1980)
6. Yang, Z.R.: Support Vector Machines for Company Failure Prediction. In: Proceedings of 2003 IEEE International Conference on Computational Intelligence for Financial Engineering, pp. 47–54 (2003)
7. Wang, X.: Corporate Financial Warning Model Based on PSO and SVM. In: 2nd International Conference on Information Engineering and Computer Science (ICIECS), pp. 1–5 (2010)
8. Bose, I., Pal, R.: Using Support Vector Machines to Evaluate Financial Fate of Dotcoms. In: Proceedings on Pacific Asia Conference on Information Systems (PACIS), Paper 42 (2005)
9. Hardle, W.K., Moro, R., Schafer, D.: Rating Companies with Support Vector Machines. DIW Berlin, Diskussionspapier, 416 (2004)
10. Hardle, W., Moro, R., Schafer, D.: Estimating Probabilities of Default With Support Vector Machines. Discussion Paper Series 2: Banking and Financial Studies. Deutsche Bundesbank, Research Centre (2008)
11. Chen, W.-H., Shih, J.-Y.: A study of Taiwan's issuer credit rating systems using support vector machines. Expert Systems with Applications 30, 427–435 (2006)
12. Chong, W., Yingjian, G., Dong, W.: Study on Capital Risk Assessment Model of Real Estate Enterprises Based on Support Vector Machines and Fuzzy Integral. In: Control and Decision Conference, pp. 2317–2320 (2008)
13. Lai, K.K., Yu, L., Wang, S.-Y., Huang, W.: An Intelligent CRM System for Identifying High-Risk Customers: An Ensemble Data Mining Approach. In: Shi, Y., van Albada, G.D., Dongarra, J., Sloot, P.M.A. (eds.) ICCS 2007. LNCS, vol. 4488, pp. 486–489. Springer, Heidelberg (2007)

14. Lai, K.K., Zhou, L., Yu, L.: A Two-Phase Model Based on SVM and Conjoint Analysis for Credit Scoring. In: Shi, Y., van Albada, G.D., Dongarra, J., Sloot, P.M.A. (eds.) ICCS 2007, Part II. LNCS, vol. 4488, pp. 494–498. Springer, Heidelberg (2007)
15. Zhou, L., Lai, K.K.: Weighted LS-SVM Credit Scoring Models with AUC Maximization by Direct Search. In: Proceedings of 2009 International Joint Conference on Computational Sciences and Optimization, pp. 7–11 (2009)
16. Zhou, L., Lai, K.K.: Multi-Agent Ensemble Models Based on Weighted Least Square SVM for Credit Risk Assessment. In: Proceedings of 2009 WRI Global Congress on Intelligent Systems, pp. 559–563 (2009)
17. Zhou, L., Lai, K.K., Yu, L.: Least squares support vector machines ensemble models for credit scoring. Expert Systems with Applications 37, 127–133 (2010)
18. Van Gestel, T., Baesens, B., Garcia, I.J.: A support vector machine approach to credit scoring (2003), `http://www.defaultrisk.com/pp_score_25.htm`
19. Van Gestel, T., Baesens, B., Suykens, J.A.K., Espinoza, M., Baestaens, D.E., Vanthienen, J., De Moor, B.: Bankruptcy prediction with least squares support vector machine classifiers. In: Proceedings of the International Conference on Computational Intelligence for Financial Engineering, CIFER, pp. 1–8 (2003)
20. Van Gestel, T., Baesens, B., Suykens, J.A.K., Van Den Poel, D., Baestaens, D.E., Willekens, M.: Bayesian kernel based classification for financial distress detection. European Journal of Operational Research 172, 979–1003 (2006)
21. Merkevičius, E., Garšva, G., Simutis, R.: Neuro-discriminate Model for the Forecasting of Changes of Companies Financial Standings on the Basis of Self-organizing Maps. In: Shi, Y., van Albada, G.D., Dongarra, J., Sloot, P.M.A. (eds.) ICCS 2007. LNCS, vol. 4488, pp. 439–446. Springer, Heidelberg (2007)
22. Danenas, P., Garsva, G.: Support Vector Machines and their Application In Credit Risk Evaluation Process. Transformations in Business & Economics 8(3(18)), 46–58 (2009)
23. Danenas, P., Garsva, G.: Credit Risk Evaluation Using SVM-Based Classifier. In: Abramowicz, W., Tolksdorf, R., Węcel, K. (eds.) BIS 2010, Part 1. LNBIP, vol. 57, pp. 7–12. Springer, Heidelberg (2010)
24. Danenas, P., Garsva, G., Gudas, S.: Credit Risk Evaluation Model Development Using Support Vector Based Classifiers. In: Proceedings of the International Conference on Computational Science (ICCS 2011), Procedia Computer Science, vol. 4, pp. 1699–1707 (2011)
25. Fan, R., Chang, K., Hsieh, C., Wang, X., Lin, C.: LIBLINEAR: A library for large linear classification. The Journal of Machine Learning Research 9, 1871–1874 (2008)
26. Zmijewski, M.: Methodological Issues Related to the Estimation of Financial Distress Prediction Models. Journal of Accounting Research 22, 59–82 (1984)
27. Danenas, P., Garsva, G., Simutis, R.: Development of Discriminant Analysis and Majority-Voting Based Credit Risk Assessment Classifier. In: Proceedings of the 2011 International Conference on Artificial Intelligence, ICAI 2011, Las Vegas, Nevada, JAV, July 23-27, vol. 1, pp. 204–209 (2011)
28. Weka 3: Data Mining Software in Java, `http://www.cs.waikato.ac.nz/ml/weka/`
29. Hall, M.A.: Correlation-based Feature Subset Selection for Machine Learning, Hamilton, New Zealand (1998)
30. UCLA-LoPucki Bankruptcy Research Database, `http://lopucki.law.ucla.edu/index.htm`