Predicting Firms' Credit Ratings Using Ensembles of Artificial Immune Systems and Machine Learning – An Over-Sampling Approach

Petr Hájek and Vladimír Olej

Institute of System Engineering and Informatics
Faculty of Economics and Administration, University of Pardubice, Studentská 84
532 10 Pardubice, Czech Republic
{petr.hajek,vladimir.olej}@upce.cz

Abstract. This paper examines the classification performance of artificial immune systems on the one hand and machine learning and neural networks on the other hand on the problem of forecasting credit ratings of firms. The problem is realized as a two-class problem, for investment and non-investment rating grades. The dataset is usually imbalanced in credit rating predictions. We address the issue by over-sampling the minority class in the training dataset. The experimental results show that this approach leads to significantly higher classification accuracy. Additionally, the use of the ensembles of classifiers makes the prediction even more accurate.

Keywords: Credit rating, artificial immune systems, machine learning, neural networks, classification performance, balanced and imbalanced dataset, SMOTE, AdaBoost.

1 Introduction

Credit rating addresses an issuer's overall capacity and willingness to meet its financial obligations, thus reducing the information asymmetry between issuers and investors. However, credit ratings are based on costly analysis performed by professionals, which is why credit rating forecasting has attracted considerable recent interest. Various artificial intelligence (AI) methods have been applied to model the complex non-linear relations between input variables and target classes (see [7] for an example of a review). Recent efforts have shown that approaches integrating feature selection process and an appropriate AI method provide the best classification performance [6]. However, although the ensembles of classifiers have shown promising results in related fields such as credit risk and bankruptcy forecasting [15], significantly insufficient attention has been paid to them in credit rating forecasting. In addition, no research has been found that examines the effect of over-sampling the minority class in credit rating data. Therefore, the aim of this paper is to examine the effect of: (1) over-sampling the minority class in credit rating data; and (2) ensembling base classifiers. The difficulty in predicting credit ratings also stems from the fact that a multitude of sources is used in the credit rating analysis,

L. Iliadis et al. (Eds.): AIAI 2014, IFIP AICT 436, pp. 29–38, 2014.

involving both quantitative data (accounting and financial data drawn from financial statements) and qualitative assessments. The sentiment analysis of corporate annual reports has recently been used to address the issue of qualitative assessment [8]. We follow this approach and use the chosen sentiment categories in addition to corporate financial indicators as input variables.

We employ two categories of AI methods to examine the given aims, (1) artificial immune systems (AISs) and (2) machine learning (ML). AISs mimic the processes and mechanisms of biological immune systems and we specifically use Artificial Immune Recognition Systems (AIRSs) [16] and the Clonal Selection Classification Algorithm (CSCA) [1]. Out of the second category, we use C4.5 decision trees (DTs) [12], Support Vector Machines (SVMs) [13], Radial Basis Function Neural Networks (RBFs) [9] and Multilayer Perceptrons (MLPs) [9]. The research to date has tended to focus on the ensembles of classifiers of the latter category while little attention has been paid to the ensembles of AISs [5].

This paper is organized as follows. In the next section, we provide the description of the dataset. Given the fact that the classes are imbalanced in the dataset, we use the Synthetic Minority Oversampling Technique (SMOTE) algorithm [10] to modify the training dataset. The third section presents the results of experiments. First, the effect of over-sampling is tested and then we employ the AdaBoost algorithm [4] to generate the ensembles of the base classifiers (AIRS1, AIRS2, AIRS2-p, CSCA; and DT, SVM, RBF, MLP). For both categories, the classification performance is examined on (imbalanced) testing dataset depending on the proportion of training dataset generated. The last section discusses the results and concludes the paper.

2 Problem Formulation and Dataset

The prediction of firms' credit ratings was realized as a two-class problem. The classes were represented by investment grade (IG, low default risk) and non-investment grade (NG, high default risk), assigned by a highly regarded Standard & Poor's rating agency in 2011. The investment grade position is critical to many investors due to the restrictions imposed on investment instruments.

We used two main groups of input variables in this study, financial indicators and sentiment indicators (see Table 1). More specifically, we used several subgroups of financial indicators such as size, profitability ratios, liquidity ratios, leverage ratios and market value ratios. Sentiment indicators refer to the business position of a company (business risk, character (reputation), organizational problems, management evaluation, accounting quality, etc.). The financial indicators were drawn from the Value Line database, while sentiment indicators were drawn from annual reports available at the U.S. Securities and Exchange Commission EDGAR System. Both groups of input variables were collected for 520 U.S. companies (selected from the Standard & Poor's database) in the year 2010, 195 classified as IG and 325 as NG. Mining and financial companies were excluded from the dataset since they require specific input variables.

The sentiment indicators required linguistic pre-processing (tokenization and lemmatization) and subsequent comparison with the financial dictionary provided by [11]. Then, the tf.idf term weighting scheme was applied to obtain the importance of terms and an average weight was calculated for each sentiment category (negative, positive, uncertainty, litigious, modal strong and modal weak). See [8] for the detailed information on the collection of data.

Table 1. Input and output variables describing the dataset

Variable		Variable	
x_1	Enterprise value	x_{11}	Dividend yield
x_2	Cash	x_{12}	Payout ratio
x_3	Revenues	x_{13}	Standard deviation of stock price
x_4	Earnings per share	x_{14}	Frequency of negative terms
x_5	Return on equity	x_{15}	Frequency of positive terms
x_6	Price to book value	x_{16}	Frequency of uncertainty terms
x_7	Enterprise value/earnings	x_{17}	Frequency of litigious terms
x_8	Price to earnings per share	x_{18}	Frequency of strong modal terms
x_9	Market debt / total capital	x_{19}	Frequency of weak modal terms
x_{10}	High to low stock price	class	{IG, NG}

The given dataset was randomly divided into training and testing dataset (2:1). This procedure was repeated five times. Thus, the training dataset O_{train} contained 347 companies, 217 classified as IG and 130 as NG. The testing dataset O_{test} covered 173 companies, 108 as IG and 65 as NG. Both datasets were imbalanced, with the less frequent NG category (minority class). There are several approaches to handle this issue. The under-sampling of the majority class may represent a good way to reduce the sensitivity of classifiers, but in our case this approach resulted in a decrease in classification performance. This may be related to both the small size of the dataset and important decision information stored in most of the objects in the majority class. Another way is to apply the over-sampling of the minority class so that all classes are represented equally in the training dataset. We used the SMOTE procedure [10] to generate additional objects (firms) for the NG class to make the training dataset balanced. Thus, O_{train} contained 433 (O_{train}^{100}) companies, 217 classified as IG and 216 as NG. These training datasets O_{train}^{100} were considered to be the base balanced training datasets with 100 % of companies. We further examined the effect of generating additional training datasets, increasing the number by 25 % up to 300 % (O_{train}^{125}, O_{train}^{150}, ... ,O_{train}^{300}). The testing set remained imbalanced and fixed for all training sets (O_{train}^{100}, O_{train}^{125}, O_{train}^{150}, ... ,O_{train}^{300}).

3 Modelling Financial and Sentiment Indicators

In this section we employed commonly used artificial immune classification algorithms, AIRS1, AIRS2, AIRS2-p and CSCA. Further, the AISs were compared with DTs, SVMs, RBFs and MLPs. We used 10-fold cross-validation on training data to find the optimum settings of the classifiers' parameters.

The measures of classification performance are represented by the averages of standard statistics applied in classification tasks [14]: true positives (TP rate), false positives (FP rate), precision (Pre) and recall (Re), F-measure (F-m), the area under the receiver operating characteristic (ROC) curve and misclassification cost (MC). F-m is the weighted harmonic mean of Pre and Re, or the Matthews correlation coefficient, which is a geometric mean of the chance-corrected variants. A ROC is a graphical plot which illustrates the performance of a binary classifier system, which represents a standard technique for summarization classifier performance over a range of tradeoffs between TP and FP error rates. In particular, FP rate is important in this study owing to its possible serious financial consequences. This is due to the significant difference between default rates of IG and NG firms. For example, Standard & Poor's reported 0.03 % for IG and 1.71 % for NG in 2011. Therefore, we designed an MC matrix, where MC = 0.03 was assigned to each false classified IG firm and MC = 1.71 to each false classified NG firm, respectively.

3.1 Artificial Immune Classification Algorithms

3.1.1 Methods

The group of AIRS algorithms represents AISs using populations [2,3]. The algorithms are based on the principle of Recognition Ball (RB) or Artificial Recognition Ball (ARB), which can be described as recognition areas or artificial recognition areas that combine feature vector (antibody) and vector class. The principle solves the issue of the completeness of AISs. Each antibody is surrounded (in the sense of antibody representation in the state space) by a small area called the RB, in which the antibody recognizes all antigens (training dataset). Further, the AIRS algorithms use the principle of a limited resource. Each ARB area competes for limited resource according to its stimulation level. The classification performance of the AIRS algorithms proposed by [16] depends on several user-defined parameters. We used the AIRS1, AIRS2 and AIRS2-p algorithms to predict the firms' credit ratings. The following parameters of the AIRSs were examined to obtain the best classification performance: affinity threshold scalar = {0.1,0.2, … ,0.9}, clonal rate = {1,2,4, … ,32}, hypermutation rate = {1,2, … ,10}, number of k nearest neighbors = {1,2, … ,10}, initial memory cell pool size = 50, number of instances to compute the affinity threshold = all, stimulation threshold = 0.9 and total resources = 150.

The CSCA [1] uses a fitness function to maximize classification accuracy (minimize misclassification accuracy). The performance of the CSCA depends on the following user-defined parameters: clonal scale factor = 1.0, initial population size = {10,20, … ,100}, number of nearest neighbors = {1,2, … ,10}, minimum fitness threshold = 1, number of partitions = 1 and total generations = {1,5, 10, ... ,100}.

3.1.2 Results

The best classification results for the AIRSs and CSCA simulations on testing dataset are shown in Table 2.

Here, we present the average performance measures over the five datasets. The measures are represented by the averages of standard statistics applied in classification tasks [14]. The results in Table 2 show that the AIRS2 performed best while the

CSCA performed significantly worse (using paired *t*-test on $P < 0.01$). Further, the sizes of training datasets suggest that the AIRSs require larger datasets to achieve good classification performance (see Fig. 1).

Table 2. Best results for AIRS1, AIRS2, AIRS2-p and CSCA on testing dataset

	AIRS1		AIRS2		AIRS2-p		CSCA	
O_{train}	O_{train}^{175}		O_{train}^{225}		O_{train}^{225}		O_{train}^{150}	
Accuracy [%]	83.24		**85.24**		83.27		80.92	
MC	41.15		35.95		37.89		**22.76**	
Class	IG	NG	IG	NG	IG	NG	IG	NG
TP rate	0.923	0.778	0.938	0.806	0.892	0.796	0.692	0.880
FP rate	0.222	0.077	0.194	0.062	0.204	0.108	0.120	0.308
Precision	0.714	0.944	0.744	0.956	0.725	0.925	0.776	0.826
Recall	0.923	0.778	0.938	0.806	0.892	0.796	0.692	0.880
F-m	0.805	0.853	0.830	0.874	0.800	0.856	0.732	0.852
ROC	0.850	0.850	0.872	0.872	0.844	0.844	0.786	0.786

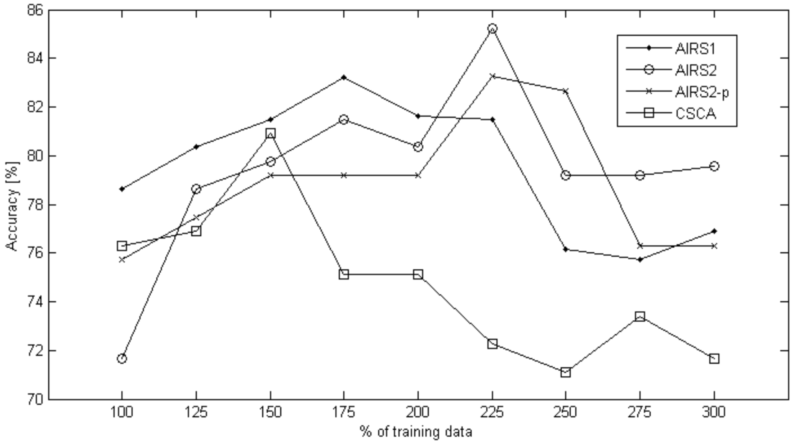

Fig. 1. Classification accuracy of AIRS1, AIRS2, AIRS2-p and CSCA on O_{test} depending on O_{train} size (Source: own)

3.2 Machine Learning and Neural Networks

3.2.1 Methods

A DT is a tree representation assigning a class to an object based on its attributes (variables), which can be continuous or discrete. An attribute with the best value of a splitting criterion is assigned to each root and intermediate node. Classes are assigned to leaf nodes in a DT. The DT with pruning was employed in this study. The classification performance of the DT depends on the following parameters: the confidence factor used for pruning (= 0.25 in this study) and the minimum number of instances per leaf = 2.

SVMs represent an essential kernel-based method with many modifications proposed recently. SVMs use kernel functions to separate the hyperplane between two classes by maximizing the margin between the closest data points. This is done in a higher-dimensional space where the data become linearly separable. We used the SVMs trained by the sequential minimal optimization (SMO) algorithm [13]. The classification performance of the SVM was tested for the following user-defined parameters: kernel functions = {polynomial, RBF}, $\gamma = 0.01$, the level of polynomial function = 2, complexity parameter $C = \{1,2,4, \dots ,256\}$, round-off error $\varepsilon = 1.0\text{E-}12$ and tolerance parameter = 0.001. The RBF was trained with the Broyden-Fletcher-Goldfarb-Shanno method. The initial centres for the Gaussian RBFs were found using a k-means algorithm. The initial sigma values were set to the maximum distance between any centre and its nearest neighbor in the set of centres. The classification performance of the RBF depends on the following user-defined parameters: maximum number iteration = not limited, minimum standard deviation for the clusters = {0.1,0.2,0.3}, the number of clusters for k-means = {1,2,4, … ,64} and ridge factor = 1.0E-8.

The MLP was trained using the backpropagation algorithm with momentum. The following parameters of the MLP were examined to achieve the best classification performance: the number of neurons in the hidden layer = {5,10,15, … ,30}, learning rate = {0.01,0.05,0.1}, momentum = 0.2 and the number of epochs = {100,500,1000}.

3.2.2 Results

The best classification results for the DT, SVM, RBF and MLP simulations on testing dataset are shown in Table 3. Here, the best classification performance was achieved by MLP (Accuracy) and DT (MC). Compared with the AISs, the MLP provided significantly better results in terms of both classification accuracy and ROCs (again, tested using paired t-test at $P < 0.01$). Furthermore, less O_{train} were required to be generated in order to learn the MLP compared with the AIRS2 (see Fig. 2). On the other hand, DT provided the lowest MC of all classifiers.

Table 3. Best results for DT, SVM, RBF and MLP algorithms on testing dataset

	DT		SVM		RBF		MLP	
O_{train}	O_{train}^{125}		O_{train}^{150}		O_{train}^{150}		O_{train}^{125}	
Accuracy [%]	86.13		87.28		84.39		**88.44**	
MC	**17.59**		30.96		36.01		24.19	
Class	IG	NG	IG	NG	IG	NG	IG	NG
TP rate	0.785	0.907	0.938	0.833	0.908	0.806	0.908	0.870
FP rate	0.093	0.215	0.167	0.062	0.194	0.092	0.130	0.092
Precision	0.836	0.875	0.772	0.957	0.738	0.935	0.808	0.940
Recall	0.785	0.907	0.938	0.833	0.908	0.806	0.908	0.870
F-m	0.810	0.891	0.847	0.891	0.814	0.866	0.855	0.904
ROC	0.871	0.871	0.886	0.886	0.884	0.881	0.947	0.947

3.3 Comparison across Classifiers

In Fig. 3 we compare the ROC achieved for the balanced and imbalanced datasets. Except for the AIRS2-p and RBF, the performance of the classifiers improved significantly (at $P < 0.05$) when the O_{train} were balanced using the SMOTE (O_{train}^{100}). Depending upon the amount of over-sampling required, neighbors from k nearest neighbors are randomly chosen. In our experiments we used $k = 5$ nearest neighbors. The results of the classification of O_{test} for the original imbalanced and balanced O_{train} are shown in Table 4.

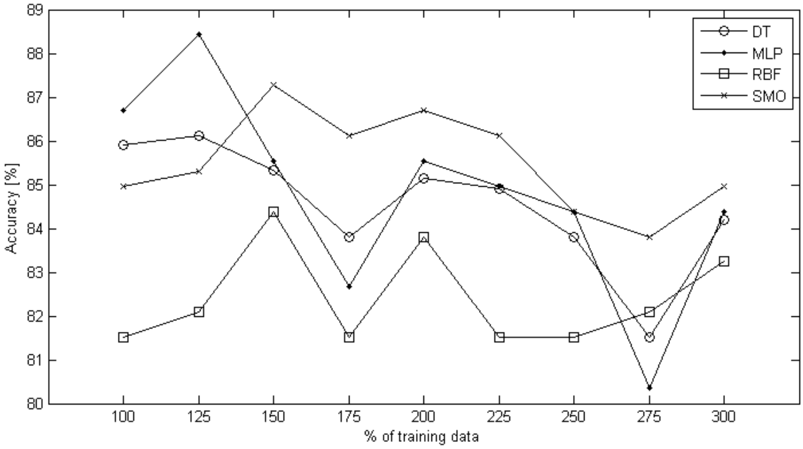

Fig. 2. Classification accuracy of DT, SVM, RBF and MLP on O_{test} depending on O_{train} size (Source: own)

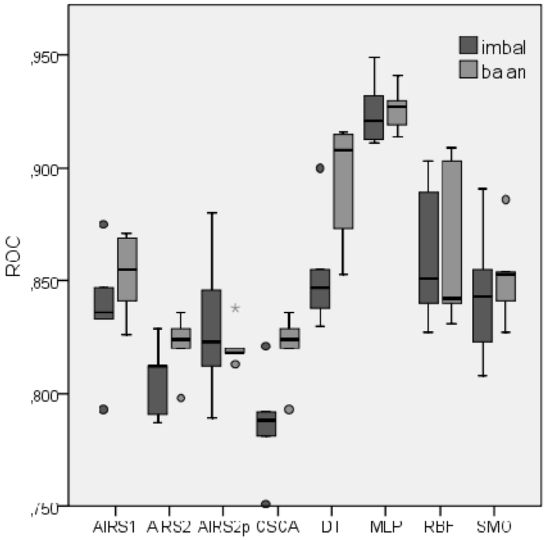

Fig. 3. ROC on testing dataset for original imbalanced and balanced O_{train} (Source: own)

We further examined the effect of the ensembles of classifiers. We employed the AdaBoost method [4] that combines many ‘weak’ classifiers to obtain an accurate learning algorithm. More precisely, AdaBoost is a meta-algorithm that can be used in conjunction with many other learning algorithms to improve their performance. This meta-algorithm is adaptive in the sense that subsequent classifiers built are tweaked in favor of those instances misclassified by previous classifiers. The number of iterations to be performed was set to 10. Again, we tested the previously described classification algorithms as base learners.

The results are depicted in Fig. 4. Here, the balanced training dataset was used from prior experiments. The ROC was significantly higher (at $P < 0.05$) for all the classifiers except for the MLP. AdaBoost provided noteworthy improvements for the DTs, AISs and SVMs especially. The results of the classification of the O_{test} for the original imbalanced and balanced O_{train} are shown in Table 5. DT with balanced data provided significantly lower MC (at $P < 0.05$) for both individual (Table 4) and ensemble classifiers (Table 5). For details, see confusion matrix in Table 6.

Table 4. Classification performance on O_{test} for the original imbalanced (i) and balanced (b) O_{train} (weighted average for IG and NG)

		AIRS1	AIRS2	AIRS2-p	CSCA	DT	SVM	RBF	MLP
TP	i	0.821	0.814	0.834	0.814	0.861	0.851	0.799	0.845
	b	0.853	0.801	0.798	0.801	0.875	0.853	0.786	0.834
FP	i	0.143	0.201	0.174	0.241	0.138	0.163	0.202	0.161
	b	0.148	0.160	0.155	0.160	0.133	0.148	0.196	0.148
F-m	i	0.822	0.814	0.834	0.810	0.862	0.851	0.800	0.846
	b	0.854	0.804	0.801	0.804	0.876	0.854	0.788	0.835
MC	i	42.42	45.99	50.80	21.56	21.29	27.99	44.80	36.14
	b	26.96	30.54	29.10	19.80	**17.89**	24.00	37.33	27.33

Table 5. Classification performance of the O_{test} for the original imbalanced (i) and balanced (b) O_{train} (weighted average for IG and NG) with AdaBoost

		AIRS1	AIRS2	AIRS2-p	CSCA	DT	SVM	RBF	MLP
TP	i	0.839	0.845	0.850	0.819	0.848	0.846	0.828	0.852
	b	0.845	0.836	0.858	0.809	0.872	0.831	0.798	0.836
FP	i	0.171	0.166	0.144	0.233	0.155	0.141	0.180	0.150
	b	0.141	0.147	0.138	0.191	0.134	0.140	0.179	0.138
F-m	i	0.840	0.857	0.852	0.816	0.849	0.848	0.829	0.853
	b	0.846	0.837	0.860	0.803	0.872	0.833	0.800	0.838
MC	i	32.70	35.14	26.97	38.99	22.51	39.52	44.09	37.83
	b	27.03	25.29	28.97	19.04	**17.28**	32.04	35.26	26.93

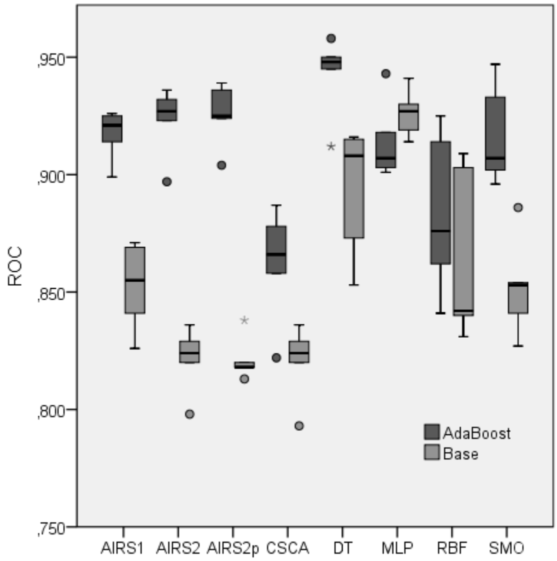

Fig. 4. ROC for base classifiers and AdaBoost (Source: own)

Table 6. Confusion matrix of the O_{test} for the balanced O_{train} trained using DT with AdaBoost

	IG	NG
IG	56.8±6.3	9.2±5.2
NG	13.0±6.1	95.0±6.1

4 Conclusion

The aim of this paper was to examine the performance of AISs and ML on a complex classification task where financial indicators are combined with sentiment analysis. We attempted to address an important issue of imbalanced dataset. On the one hand, the over-sampling of the minority class showed promising classification improvement, yet on the other hand, the under-sampling of the majority class resulted in weaker classification performance. We performed many experiments to find the best setting of the learning parameters of individual classifiers to achieve the best classification performance. The results also confirmed that an ensemble of weaker base classifiers may perform better than the individual classifiers, especially in terms of ROC measure. More importantly, the decrease in MC may lead to the substantial financial savings of investors. For example, the ensemble of DTs trained on balanced data reduced the MC from 21.29 to 17.28 (18.8% savings). Thus, this study represents a good basis for further experiments in the field of financial distress forecasting, where the problem of imbalanced datasets is usually to be addressed. Additionally, we encourage future research in multi-class datasets.

The experiments in this study were carried out in Statistica 10 (linguistic preprocessing) and Weka 3.7.5 (Artificial Immune Classification Algorithms and ML) in MS Windows 7 operating system.

Acknowledgments. This work was supported by the scientific research project of the Czech Sciences Foundation Grant No: 13-10331S.

References

1. Brownlee, J.: Clonal Selection Theory Algorithm and CLONALG: The Clonal Selection Classification Algorithm (CSCA). Victoria Australia, Centre for Intelligent Systems and Complex Processes, Swinburne University of Technology, Technical Report ID: 2-01 (2005)
2. Dasgupta, D.: Artificial Immune Systems and their Applications. Springer, Berlin (1999)
3. De Castro, L.N., Timmis, J.: Artificial Immune Systems: A New Computational Intelligence Approach. Springer, Berlin (2002)
4. Freund, Y., Schapire, R.E.: Experiments with a New Boosting Algorithm. In: 13th International Conference on Machine Learning, San Francisco, pp. 148–156 (1996)
5. García-Pedrajas, N., Fyfe, C.: Construction of Classifier Ensembles by means of Artificial Immune Systems. Journal of Heuristics 14(3), 285–310 (2008)
6. Hajek, P.: Municipal Credit Rating Modelling by Neural Networks. Decision Support Systems 51(1), 108–118 (2011)
7. Hajek, P., Olej, V.: Credit Rating Modelling by Kernel-Based Approaches with Supervised and Semi-Supervised Learning. Neural Computing and Applications 20(6), 761–773 (2011)
8. Hájek, P., Olej, V.: Evaluating Sentiment in Annual Reports for Financial Distress Prediction Using Neural Networks and Support Vector Machines. In: van Zee, G.A., van de Vorst, J.G.G. (eds.) Shell Conference 1988. LNCS, vol. 384, pp. 1–10. Springer, Heidelberg (1989)
9. Haykin, S.: Neural Networks: A Comprehensive Foundation, 2nd edn. Prentice-Hall Inc., New Jersey (1999)
10. Chawla, N.V., Bowyer, K.W., et al.: SMOTE: Synthetic Minority Over-sampling Technique. Journal of Artificial Intelligence Research 16, 321–357 (2002)
11. Loughran, T., McDonald, B.: When Is a Liability Not a Liability? Textual Analysis, Dictionaries, and 10-Ks. The Journal of Finance 66(1), 35–65 (2011)
12. Quinlan, J.R.: C4. 5: Programs for Machine Learning. Morgan Kaufmann (1993)
13. Platt, J.C.: Fast Training of Support Vector Machines using Sequential Minimal Optimization. In: Schoelkopf, B., Burges, C., Smola, A. (eds.) Advances in Kernel Methods - Support Vector Learning. MIT Press (1998)
14. Powers, D.M.W.: Evaluation: from Precision, Recall and F-measure to ROC, Informedness, Markedness and Correlation. Journal of Machine Learning Technologies 1(2), 37–63 (2011)
15. Ravi Kumar, P., Ravi, V.: Bankruptcy Prediction in Banks and Firms via Statistical and Intelligent Techniques - A Review. European Journal of Operational Research 180(1), 1–28 (2007)
16. Watkins, B.A., Timmis, J., Boggess, L.: Artificial Immune Recognition System (AIRS): An Immune Inspired Supervised Learning Algorithm. Genetic Programming and Evolvable Machines 5, 291–317 (2004)