

Credit risk prediction using support vector machines

Jan-Henning Trusterff · Paul Markus Konrad · Jens Leker

Published online: 24 July 2010
© Springer Science+Business Media, LLC 2010

Abstract The main purpose of this paper is to examine the relative performance between least-squares support vector machines and logistic regression models for default classification and default probability estimation. The financial ratios from a data set of more than 78,000 financial statements from 2000 to 2006 are used as default indicators. The main focus of this paper is on the influence of small training samples and high variance of the financial input data and the classification performance measured by the area under the receiver operating characteristic. The resolution and the reliability of the predicted default probabilities are evaluated by decompositions of the Brier score. It is shown that support vector machines significantly outperform logistic regression models, particularly under the condition of small training samples and high variance of the input data.

Keywords Support vector machines · Credit risk prediction · Default classification · Estimation of probabilities of default · Training sample size · Accounting data

JEL Classification C14 · G33

1 Introduction

There is a wide range of quantitative methods to assess the creditworthiness of loan applicants and to estimate probabilities of default (PD). As well-developed statistical models often outperform a subjective credit risk assessment (Sun 2007), quantitative methods are common in banks' credit risk assessment. The most common parametric techniques are linear and quadratic discriminant analysis as well as generalized linear models such as logistic regression models (LRM). Besides these statistical-driven techniques, different non-parametric methodologies such as decision tree learning, genetic algorithms or neuronal networks can be distinguished (Chen and Chiou 1999;

J.-H. Trusterff (✉) · P. M. Konrad · J. Leker
Institute of Business Administration, University of Münster, Leonardo-Campus 1,
48149 Münster, Germany
e-mail: trusterff@uni-muenster.de

Yobas et al. 2000). The latter are often superior to parametric methods (Atiya 2001; Varetto 1998), but suffer from non-convex optimization problems. Apart from optimization problems, the indicator weighting often contradicts the economic default hypothesis associated with the financial ratios used as predictor variables. Other enhanced credit risk models take macroeconomic conditions in addition to firm-specific characteristics into account (Carling et al 2007; Butera and Faff 2006). The dynamic of macroeconomic conditions and a time lagged impact on firm-specific characteristics lead to a high complexity of these models.

Furthermore, all upper mentioned approaches are designed to specify the classifier by minimizing the classification error on training data. Schölkopf and Smola (2002) point out that empirical risk minimization does not imply low structural risk, i.e. a small classification error on unseen test data. Thus, regularization methods or outlier adjustments are necessary to ensure both unbiased classification and high generalization.

Support Vector Machines (SVM) are derived from the statistical learning theory and follow a structural risk minimization principle as shown by Boser et al. (1992) and Cortes and Vapnik (1995). To obtain classifiers, these powerful learning systems merge efficient algorithms from the optimization theory and elements of the statistical learning theory. Huang et al. (2004) underline the intersectional character of SVM by combining the advantages of theory-driven methods and data-driven machine learning methods. The basic idea of SVM is to define a hyperplane that geometrically separates binary classes. As demonstrated by Vapnik (2000), the optimal hyperplane is obtained by maximizing the margin between the data points of the two classes whereby a structural risk minimum is achieved. Nonlinear SVM classifiers apply kernel functions to map the data from the input space into a higher dimensional feature space, where an optimal separating hyperplane may be constructed (Schölkopf and Smola 2002). In contrast to other nonlinear classification methodologies like neural networks, the convexity of the underlying optimization problem is ensured.

The classification performance is often superior to other artificial intelligence approaches (e.g. neural networks or genetic algorithms) and classical statistical techniques (Ravi Kumar and Ravi 2007; Cristianini and Shawe-Taylor 2006; Abe 2005). This has been shown, for example, in bioinformatic applications, text categorization, e-mail spam detection, and general pattern recognition problems (Baesens et al. 2003). The central advantages of SVM include the geometrical representation of the classifier as a hyperplane, the formulation of a convex optimization problem that leads to a unique optimum, and the possibility to estimate an upper bound of the generalization error on unseen test data. The resulting classifier generalizes well even in high dimensional input spaces and under small training sample conditions. As the amount of available financial default data is typically small and often of low quality, this classifier seems particularly suitable for credit risk estimation, including default classification (Härdle et al. 2005) and PD estimation.

This paper gives a comprehensive introduction to SVM and compares the credit rating capabilities of a SVM classifier to a LRM regarding default classification and PD prediction. Chen et al. (2006) and Härdle et al. (2007) have shown a superior classification performance of the SVM. Little availability of accounting data from bankrupt companies and high variance of the derived financial ratios is a typical problem of credit risk prediction. Therefore, we particularly focus on the influence of high variance of the training data and small training sample conditions. Moreover, the empirical analysis is based on a large up-to-date data sample and in contrast to the aforementioned publications adequate statistical measurements are considered to evaluate classification as well as accuracy and reliability of the predicted probabilities of default.

The paper is structured as follows: Section 2 provides an introduction into the theory of SVM.¹ Section 3 describes the empirical research design including data set composition and the financial ratios used for default estimation. The results of default classification accuracy and goodness of PD prediction are presented in Sect. 4, followed by a conclusion in Sect. 5.

2 SVM classifier for credit scoring

Consider a given training data set $\{(\mathbf{x}_k, y_k)\}_{k=1}^N$ with input data $\mathbf{x}_k \in \mathbb{R}^n$ and dependent binary class labels $y_k \in \{-1, 1\}$, where default is denoted as $y_k = 1$ and non-default is denoted as $y_k = -1$. The input data \mathbf{x}_k represents the observation of the n applied default indicator variables for company k .

According to Vapnik (1998), Hastie et al. (2001) and Schölkopf and Smola (2002) a separating hyperplane $\mathbf{w}^T \varphi(\mathbf{x}) + b = 0$ satisfies for a non-linear and separable case

$$\mathbf{w}^T \varphi(\mathbf{x}_k) + b \geq +1 \quad \text{if } y_k = 1 \tag{1}$$

$$\mathbf{w}^T \varphi(\mathbf{x}_k) + b \leq -1 \quad \text{if } y_k = -1 \tag{2}$$

equivalent to

$$y_k [\mathbf{w}^T \varphi(\mathbf{x}_k) + b] \geq 1 \tag{3}$$

with the weight vector \mathbf{w}^T and the bias b . The resulting SVM classifier is

$$y(\mathbf{x}) = \text{sgn}(\mathbf{w}^T \varphi(\mathbf{x}) + b). \tag{4}$$

The function $\varphi(\mathbf{x}) : X \rightarrow F$ is a non-linear map from the input space X to some high dimensional feature space F , where a linear separation is possible (Cristianini and Shawe-Taylor 2006). However, a perfect non-linear separation is not desirable in practical use, as it often leads to a low generalization ability.

For a non-separable case, it is necessary to introduce a slack variable ξ with $\xi_k \geq 0 \forall k$ to tolerate misclassification in the set of margin constraints. Thus, constraint (3) still holds for overlapping classes. Misclassification only occurs when $\xi_k > 1$. The classification constraint (3) follows as

$$y_k [\mathbf{w}^T \varphi(\mathbf{x}_k) + b] \geq 1 - \xi_k. \tag{5}$$

The decision function that classifies test data points results from the solution of the convex optimization problem

$$\max_{\mathbf{w}, b, \xi} C \Rightarrow \min_{\mathbf{w}, b, \xi} \frac{1}{2} \|\mathbf{w}\|^2 \tag{6}$$

$$\text{subject to } y_k [\mathbf{w}^T \varphi(\mathbf{x}_k) + b] \geq 1 - \xi_k, \tag{7}$$

$$\xi_k \geq 0, k = 1, \dots, N \tag{8}$$

The equivalent objective function for the optimization problem (6) is considered as

¹ Regarding the statistical backgrounds of logistic regression models, please refer to Hastie and Tibshirani (1990), Hosmer and Lemeshow (2000) and McLachlan (2004).

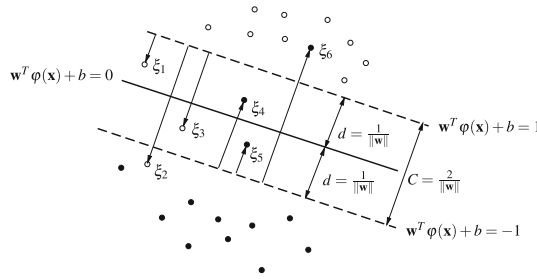


Fig. 1 Soft-margin SVM classifier for a non-separable case of overlapping classes. The *solid line* shows the separating hyperplane that is a linear decision boundary in the mapped feature space F . The maximal margin of width $C = \frac{2}{\|w\|}$ is represented by *broken lines*. ξ_k is the distance of data points x_k from the correct side of the maximal margin

$$\min_{w,b,\xi} \mathcal{F}_P(w, \xi) = \frac{1}{2} w^T w + \gamma \sum_{k=1}^N \xi_k \tag{9}$$

$$\text{subject to } y_k [w^T \varphi(x_k) + b] \geq 1 - \xi_k, \tag{10}$$

$$\xi_k \geq 0, k = 1, \dots, N \tag{11}$$

where $\frac{1}{2} w^T w$ maximizes the margin between the two classes and the separating hyperplane, while $\gamma \sum_{k=1}^N \xi_k$ minimizes the classification error.

As illustrated in Fig. 1, the so-called soft-margin classifier is obtained by maximizing the margin C subject to the classification error. The classification error $\sum_{k=1}^N \xi_k$ and the classifier capacity have to be controlled to ensure high generalization on unknown test data. The capacity is controlled via $\frac{1}{2} w^T w$ (Schölkopf and Smola 2002). The higher the capacity, the more suitable is the selected classifier to separate the training data. However, as shown by Vapnik (1998) this does not guarantee a small test error and even increases the uncertainty of the test error. The tuning parameter γ is a positive real number that weights the influence of misclassification on the objective function $\mathcal{F}_P(w, \xi)$. The (primal) Lagrangian is considered as

$$\mathcal{L}(w, b, \xi; \alpha, \nu) = \mathcal{F}(w, \xi) - \sum_{k=1}^N \alpha_k (y_k [w^T \varphi(x_k) + b] - 1 + \xi_k) - \sum_{k=1}^N \nu_k \xi_k \tag{12}$$

with Lagrange multipliers $\alpha_k > 0, \nu_k > 0 \forall k$. The solution function $f(x)$ can be written as

$$f(x) = w^T \varphi(x) + b \tag{13}$$

$$= \sum_{k=1}^N \alpha_k y_k \varphi^T(x) \varphi(x_k) + b \tag{14}$$

According to Vapnik (1998), Boser et al. (1992) no explicit construction of $\varphi(x)$ is necessary. The solution function can be computed just by using inner products between a particular transformed test data point $\varphi(x)$ and all transformed training data points $\varphi(x_k)$. This computation in a huge (even infinite) dimensional feature space F is possible by application of a kernel function $K(x, z)$, such that for all $x, z \in X$

$$K(x, z) = \varphi^T(x) \varphi(z) \tag{15}$$

so that (14) follows as

$$f(\mathbf{x}) = \sum_{k=1}^N \alpha_k y_k K(\mathbf{x}, \mathbf{x}_k) + b. \quad (16)$$

The kernel function $K(\mathbf{x}, \mathbf{x}_k)$ returns a real valued number characterizing the similarity of a particular test data point \mathbf{x} and each training data point \mathbf{x}_k (Schölkopf and Smola 2002). As the default behavior of the training data and its similarity to the test data is known, a default prediction for the test data can be derived. In this paper, a radial basis kernel function

$$K(\mathbf{x}, \mathbf{x}_k) = \exp\left(-\frac{\|\mathbf{x} - \mathbf{x}_k\|^2}{\sigma^2}\right) \quad (17)$$

is implemented. This kernel function is a standard approach (Evgeniou et al. 2000) and achieves best accuracies on different benchmark data sets (van Gestel et al. 2004). However, the improvement of the statistical fitting leads to an intransparent classifier response, as the non-monotonic prediction of a nonlinear SVM may imply a non intuitive marginal impact of each default indicator that is in contrast to its underlying financial default hypothesis.

Suykens and Vandewalle (1999) propose a least-squares SVM (LS-SVM) classifier with a least-squares cost function and replace the inequality constraints by equality constraints. Hence (9) to (11) are modified into

$$\min_{\mathbf{w}, b, e} \mathcal{F}_P(\mathbf{w}, e) = \frac{1}{2} \mathbf{w}^T \mathbf{w} + \gamma \frac{1}{2} \sum_{k=1}^N e_k^2 \quad (18)$$

$$\text{subject to } y_k [\mathbf{w}^T \varphi(\mathbf{x}_k) + b] = 1 - e_k \quad \forall k \quad (19)$$

where e_k plays a similar role as the slack variable ζ_k in the soft-margin SVM formulation (5). The LS-SVM simplify the solving of the optimization problem because of equality constraints and a linear Karush-Kuhn-Tucker system (Suykens et al. 2002). As computation is accelerated and performance comparable to other SVM formulations, we will use a LS-SVM classifier in the following analysis.

3 Data and methods

The data sample contains financial reports of 31,049 German companies from 2000 to 2006, thereof 1,112 companies had a default or failure to pay their credit obligations. The underlying financial reports are prepared in accordance with the German Accounting Standards (HGB) or the International Financial Reporting Standards (IFRS).

To ensure highest data quality, all financial ratios were calculated on the basis of verified balance sheets and income statements. The financial data were verified by comparing the published superordinated items of the balance sheet and income statement to checksums that were derived by aggregating the corresponding subordinated items. If these checksums exceeded the value of the published superordinated item by more than EUR 100, the corresponding financial data record was excluded from the study.

A particular financial data record of a company was labeled as default, if a credit event occurred within 2 years from the corresponding reporting date. The credit event was

Table 1 Composition of total data samples over time horizon

Year	2000	2001	2002	2003	2004	2005	2006	Total
Non-defaults	9,002	11,847	13,958	14,907	15,866	11,884	657	78,122
Defaults	4	109	416	333	309	86	1	1,259
Total	9,007	11,955	14,374	15,240	16,176	11,970	659	79,381

Source: Authors calculation. The number of plausible financial reports of defaulted and non-defaulted companies are given for the total data sample (Total). The little number of available default reports in 2006 results from time lag effects, because the data is based on a default labeling in March 2007. The default label could only be assigned to companies, which defaulted before this date

defined in accordance with the Basel Committee on Banking Supervision (2006). Thus, an obligor is denoted as default, if the payment of credit obligations is unlikely or past due more than 90 days. The default denotation, subject to the default horizon and default definition, is crucial for a comparison of performance studies on default classification and PD estimation.

The data was prepared as follows. In a first step the total data sample, including financial data and associated default information, was randomly partitioned into a training sample and a hold-out test sample. To avoid a biased training of the classifier, which might be due to external and non-financial variables such as industry, company size, report date, or legal form, we carefully prepared the training data. Each default data record was matched to a similar non-default data record (one-to-one-matching) regarding legal form, industry sector, date of the financial report, and company size.² Thus, the explanatory default information inherent in the training data is balanced.

The classifier was estimated on this matched training sample and tested on the non-matched test sample to evaluate its generalization ability. The data samples often contained several financial reports of the same company referring to different report dates. To avoid a bias due to different length or irregularity of time series, the training and test samples were composed as company-unique data samples.³

In a second step, the total data sample was modified into a company-unique sample, that only contains the most recent financial data record of each default and non-default company. This representative sample of the global company portfolio, denoted as calibration sample, was used to estimate calibrated PD. The number of companies and financial reports available for the analysis are displayed in Table 1.

Classification and prediction of PD may be affected by a high variance of the data and a varying partitioning into training and test sample. To analyze the impact of variance and partitioning on classification and PD prediction, both was evaluated for alternative outlier adjustments and different training/test partitions (Table 2).

The outlier adjustment was implemented as a simple percentile-based elimination of outlier data. A data record (y_k, \mathbf{x}_k) was considered as outlier and excluded from the corresponding data sample, if

$$\mathbf{x}_k^{ND} < P_\epsilon^{ND}(\mathbf{x}^{ND}) \vee \mathbf{x}_k^{ND} > P_{1-\epsilon}^{ND}(\mathbf{x}^{ND}) \quad (20)$$

² Regarding company size, three classes of sales volume are distinguished: small (<EUR 8 million), medium (EUR 8–16 million) and large (>EUR 16 million). These sales classes refer to §267 HGB (German Accounting Standard).

³ A company-unique data sample includes N data points \mathbf{x}_k of the companies $k = 1, \dots, N$.

Table 2 Composition of data samples

	Total	Training	Training*	Testing	Calibration
Non-default					
Reports	78,122	58,627	823	7,492	29,937
Companies	29,937	22,445	823	7,492	29,937
Default					
Reports	1,259	960	841	271	1,112
Companies	1,112	841	841	271	1,112
Total					
Reports	79,381	59,587	1,664	7,763	31,049
Companies	31,049	23,286	1,664	7,763	31,049

Source: Authors calculation. The number of reports and companies are given for the total sample (Total), that is randomly partitioned into training data (Training) and hold-out test data (Testing). Here, 75% of the total data was allocated for training and 25% for testing. The training data was used to compose a matched training sample (Training*), where the included default and non-default companies are similar regarding industry, company size, report date and legal form. The calibration data (Calibration) is a company-unique data sample that was used to estimate calibrated probabilities of default

and

$$\mathbf{x}_k^D < P_\epsilon^D(\mathbf{x}^D) \vee \mathbf{x}_k^D > P_{1-\epsilon}^D(\mathbf{x}^D), \quad (21)$$

where \mathbf{x}_k^D (\mathbf{x}_k^{ND}) is the input data of the particular default (non-default) data record and P_ϵ^D (P_ϵ^{ND}) is the $\epsilon\%$ percentile of the default (non-default) data. The higher the adjustment level ϵ , the more restrictive is the outlier adjustment and the lower is the variance of default and non-default data.

To analyze the impact of outliers on the classification of unseen test data, the training data was adjusted for outliers at different levels. The PD estimation may be affected by a biased classifier estimation and a biased calibration of the PD prediction. Therefore, the accuracy of the PD prediction was evaluated separately for adjusted training data and adjusted calibration data.

Both the SVM classifier and the LRM include the same financial ratios as default indicators to ensure a methodical performance evaluation. The considered default indicators were selected based on their univariate discrimination power measured by the area under the receiver operating characteristic curve.

As a large number of default explaining indicators allows only little performance gains due to mutual correlation and leads to the problem of the so-called curse of dimensionality, the number of applied variables has to be reduced (Theodoridis and Koutroumbas 2006). Therefore, the indicators were ranked according to their univariate discrimination power and the low performing ones were gradually eliminated until the Pearson correlation coefficient of the remaining financial ratios did not exceed 0.45. It should be noted that the Pearson correlation coefficient does not indicate complex non-linear relationships. These dependencies were not further considered because an approach that considers the wide universe of possible non-linear relationships might lead to the finding of spurious relations among the explaining variables. Hence, our approach might not capture all dependencies between the explaining variables but the critical effects of dependencies on model specification are accounted for.

Finally the following 5 indicators were chosen from a catalogue of original 19 financial ratios to model the credit risk of medium and large sized companies of the manufacturing, trading and service industry.

1. Cash flow return on investment:	$CFROI = \frac{\text{Cash flow}}{\text{Total assets}}$
2. Net debt ratio:	$NDR = \frac{\text{Financial liabilities} - \text{Cash}}{\text{Capital employed}}$
3. Interest coverage:	$IC = \frac{\text{EBITDA}}{\text{Interest expenses}}$
4. Equity ratio:	$ER = \frac{\text{Equity}}{\text{Total assets}}$
5. Liability turnover:	$LT = \frac{\text{Accounts payable}}{\text{Net sales}}$

This set of ratios refers to profitability, capital structure and the relation of cash flow to net debt. Hence, a wide spectrum of credit risk related information of the balance sheet and the income statement is covered. The selected indicators are not appropriate for banks and insurance companies, because their financial reporting structure and business model fundamentally differs from the industries mentioned above. Therefore, financial institutions and small entities with less than EUR 5 million sales volume are not included in the data sample.

The comparison of the SVM classifier and a LRM was structured into three steps. First, the SVM classifier and the LRM were specified. Second, the classification performance of the SVM classifier and the LRM on the test sample was evaluated and compared using the test suggested by DeLong et al. (1988). Finally, the accuracy of the estimated PD of 25 equally sized rating classes was evaluated by the Brier score and its decompositions.

The probability of default $\pi(\mathbf{x})$ is provided by the LRM

$$\pi(\mathbf{x}) = \frac{\exp(\beta_0 + \beta^T \mathbf{x})}{1 + \exp(\beta_0 + \beta^T \mathbf{x})} \quad (22)$$

with the coefficient vector $\beta^T = (\beta_1, \dots, \beta_n)$ and the intercept β_0 . Coefficient vector and intercept were estimated by the maximum likelihood method implemented in the MATLAB[®] Statistics Toolbox.

The non-linear SVM classifier (16) was trained by the MATLAB[®] Bioinformatics Toolbox. The choice of the required kernel parameter σ and the margin control parameter γ is crucial. Because a global and unique optimum of the parameters cannot be obtained analytically, a range of values must be tried before an adequate parameter set can be selected (Schölkopf and Smola 2002; Cristianini and Shawe-Taylor 2006). The generalization parameter γ and the kernel parameter σ of the radial basis kernel were specified by selecting a parameter set that maximizes the classification accuracy on a randomly separated subset of the training sample. To obtain an adequate parameter set, the classification accuracy was evaluated for $\gamma = 0, \dots, 25$ and $\sigma = 1, \dots, 60$ in interval steps of 0.2. The SVM classifier was finally trained and tested with the specified adequate parameter set on the matched training sample and test sample.

A common criterion to measure classification accuracy is the area A under the receiver operating characteristic (ROC). The empirical ROC curve is a plot of the hitrate (correctly classified as default) $HR(c)$ against the corresponding false alarm rate (falsely classified as default) $FAR(c)$ for alternative cut-off values c . The cut-off values are used to classify

observed data points \mathbf{x} as default ($y = 1$) or non-default ($y = -1$). The area under the ROC curve is equal to the probability

$$A = P(S_{ND} < S_D) + \frac{1}{2}P(S_{ND} = S_D). \tag{23}$$

The rating scores of default companies S_D and non-default companies S_{ND} are continuous random variables. Their underlying distribution results from the obtained classifier (Engelmann et al. 2003; Bamber 1975).

The classification accuracy of the SVM classifier and the LRM were compared by testing the difference of the two correlated areas under the ROC according to DeLong et al. (1988). We constructed this test as a one-tailed test with the null hypothesis that the areas under the ROC are equal, against the alternative hypothesis that the classification accuracy of the SVM classifier is superior to the LRM. The compared ROC curves were checked for intersection to make sure that the observed area under the ROC curve is an overall difference of classification performance for companies with higher and lower default risk.

$$H_0 : A_{SVM} = A_{LRM} \tag{24}$$

$$H_1 : A_{SVM} > A_{LRM} \tag{25}$$

Apart from classification accuracy, a reliable PD forecasting is very important for credit risk assessment. While classification focuses on the event of default and only allows for binary class assignments, pricing and rating of credit obligations require a measure to qualify the risk exposure. A ranking of credit obligations is only possible, when the underlying variable is at least ordinally scaled. Assuming a portfolio of non-default companies only, an evaluation of classification accuracy is not reasonable. However, the reliability of probability forecasts should certainly be considered, since an accurate grading of these companies according to their individual risk exposure requires an unbiased estimation of the PD. Furthermore, the banking regulation framework of the Basel Committee on Banking Supervision (2006) stresses the reliability issue as well as the classification performance.

The LRM (22) directly provides a probability of default $\pi(\mathbf{x})$. However, due to the matching procedure, the estimated model implies an a-priori probability of default of about 50%. The necessary recalibration of the LRM and the transformation of the SVM classifier values into calibrated forecasts of PD is achieved by fitting a non-linear model

$$\hat{p}(s) = \frac{1}{1 + \alpha_1 \exp(\alpha_2 s)} \tag{26}$$

where \hat{p} is the estimated probability of default and s is the rating score provided by the LRM and the SVM classifier. The model parameters α_1 and α_2 are obtained by a non-linear fitting of the model.

The rating score s_{LRM} provided by the LRM is given by

$$s_{LRM} = \hat{\beta}_0 + \hat{\beta}^T \mathbf{x} \tag{27}$$

where $\hat{\beta}_0$ and $\hat{\beta}$ are the intercept and coefficient vector resulting from the maximum likelihood estimation of model (22). The SVM rating score s_{SVM} is given by the SVM classifier (16) and thus

$$s_{SVM} = \sum_{k=1}^N \alpha_k y_k K(\mathbf{x}, \mathbf{x}_k) + b. \tag{28}$$

The estimation of model (26) was accomplished in two steps. First, all data records of the calibration sample were ordered by their issued rating score. The ordered sample is then segmented into 25 ranked and equally sized rating classes. Finally, the model (26) was estimated by a non-linear least squares regression of the average default frequency on the average rating score of all rating classes, whereby the default frequency was employed as an estimator of the PD.

A common measure to evaluate the PD prediction is the Brier score and its decomposed components (Hersbach 2000). The Brier score is defined as

$$BS = \frac{1}{N} \sum_{i=1}^m N_i \left[o_i(1 - \hat{p}_i)^2 + (1 - o_i)\hat{p}_i^2 \right] \quad (29)$$

where o_i is the observed default frequency, \hat{p}_i is the estimated probability of default and N_i is the size of rating class $i = 1, \dots, m$. The lower the Brier score, the better is the probabilistic prediction. In the case of equally sized rating classes, formula (29) can be simplified to

$$BS = \frac{1}{m} \sum_{i=1}^m \left[o_i(1 - \hat{p}_i)^2 + (1 - o_i)\hat{p}_i^2 \right]. \quad (30)$$

The Brier score (30) can be decomposed into the uncertainty of the default frequency U , the resolution RS and the reliability RL of the PD prediction:

$$BS = RL - RS + U \quad (31)$$

The reliability

$$RL = \frac{1}{m} \sum_{i=1}^m [o_i - \hat{p}_i]^2 \quad (32)$$

is the mean squared prediction error. The lower the RL , the better is the accuracy of the PD prediction. The reliability concerning a particular rating class i can be measured by the logarithm of the relation of observed default frequency and the predicted PD

$$LRL_i = \ln(o_i) - \ln(\hat{p}_i) \quad (33)$$

that relativizes the prediction error to the estimated PD of rating class i . In this way, the prediction error is standardized and more easily comparable throughout all rating classes. As long as $o_i > 0$ holds for all $i = 1, \dots, m$, LRL_i is well defined.

The resolution

$$RS = \frac{1}{m} \sum_{i=1}^m (o_i - o)^2, \quad (34)$$

is the average squared deviation of the observed default frequencies o_i from the observed global default frequency o of the data sample. The higher the resolution, the better is the refinement and the reliability of the rating class assignment.

Assuming binomially distributed default events, the uncertainty of the default frequency U is equal to the variance of the global default frequency and given by

$$U = o(1 - o). \quad (35)$$

This last component of the Brier score is independent of the classifier and inherent in the data sample. Hence, Brier scores of classifiers always have to be compared on the basis of the same data sample.

4 Results

Table 3 shows the area under the ROC and its standard error $\hat{\sigma}(A)$ of the SVM classifiers and the LRM. The last column results from the test of the difference of the two correlated areas according to DeLong et al. (1988), assuming an outperformance of the SVM classifier. It displays the probability of a false rejection of the null hypothesis (24). Engelmann et al. (2003) showed that asymptotic normality of A is approximately given for small samples with 50 defaults. As training and test sample include more than 100 default data points, normality may be assumed and the tests of the performance differences may be considered as reliable.

The non-linear SVM classifier yields a significantly higher classification accuracy on the training and test sample for all alternative partition settings. The extent of outperformance differs from training to test sample. Regarding the classification of training data, the absolute difference of the two compared areas is nearly constant at about 7% and highly significant for all training and test partitions. However, regarding the classification of test data, this difference decreases from 6 to 4%, if the training partition is enlarged from 75 to 90% of the data sample. The smaller the training sample, the more superior is the classification performance of the SVM classifier.

The standard error $\hat{\sigma}(A)$ decreases for larger training samples and increases for smaller test samples. The change of the standard error mainly result from a varying sample size. However, the standard errors of the estimated areas of the SVM classifier are smaller than those of the LRM for all partitioning settings. Despite a worsened confidence level due to an increasing standard error and a decreasing performance difference, the outperformance of the SVM classifier is still statistically significant at the 1% level.

The classification accuracies on a non-adjusted test sample for alternative outlier adjustments of the training sample are displayed in Table 4.

Table 3 Classification accuracy of the SVM classifier and the LRM

Partition	Size (%)	SVM		LRM		p	# Default
		A	$\hat{\sigma}(A)$	A	$\hat{\sigma}(A)$		
Training	75	0.80608***	0.01073	0.73135	0.01248	0.0000	836
Test	25	0.78977***	0.01295	0.72698	0.01511	0.0000	276
Training	83	0.81233***	0.00997	0.73388	0.01176	0.0000	919
Test	17	0.79024***	0.01605	0.73802	0.01802	0.0000	193
Training	90	0.81533***	0.00948	0.74250	0.01105	0.0000	1,006
Test	10	0.77313***	0.02243	0.73129	0.02495	0.0025	106

Source: Authors calculation. The areas under the receiver operating characteristic A and standard errors $\hat{\sigma}(A)$ are given for different partitionings of training and test data. The probability p of a false rejection of the hypothesis $A_{SVM} = A_{LRM}$ in favor of the alternative hypothesis $A_{SVM} > A_{LRM}$ results from the test according to DeLong et al. (1988). Outperformance of the SVM classifier is significant at confidence level *** $p < 0.01$, ** $p < 0.05$ and * $p < 0.1$

Table 4 Classification accuracy of the SVM classifier and the LRM (test data)

Partitioning (training/test)	ε (%)	SVM		LRM		p
		A	$\hat{\sigma}(A)$	A	$\hat{\sigma}(A)$	
75%/25%	0	0.7898**	0.0129	0.7270	0.0151	0.0000
	0.5	0.7908***	0.0128	0.7718	0.0132	0.0012
	1	0.8010***	0.0125	0.7695	0.0130	0.0000
	2	0.7990***	0.0128	0.7678	0.0132	0.0000
83%/17%	0	0.7902**	0.0161	0.7380	0.0180	0.0000
	0.5	0.7970*	0.0152	0.7803	0.0156	0.0500
	1	0.8062**	0.0150	0.7796	0.0151	0.0015
	2	0.7964**	0.0153	0.7764	0.0156	0.0771
90%/10%	0	0.7731***	0.0224	0.7313	0.0250	0.0025
	0.5	0.7771	0.0221	0.7769	0.0208	0.9591
	1	0.7784	0.0221	0.7722	0.0206	0.5433
	2	0.7761	0.0228	0.7612	0.0222	0.2004

Source: Authors calculation. The areas under the receiver operating characteristic A and standard errors $\hat{\sigma}(A)$ are given for different test partitions and a varying outlier adjustment of the training data at level ε . The probability (p) of a false rejection of the hypothesis $A_{SVM} = A_{LRM}$ in favor of the alternative hypothesis $A_{SVM} > A_{LRM}$ results from the test according to DeLong et al. (1988). Outperformance of the SVM classifier is significant at confidence level *** $p < 0.01$, ** $p < 0.05$ and * $p < 0.1$

While the classification performance of the SVM classifier on the test data is barely influenced, the classification performance of the LRM significantly increases as outliers are removed from the training data. In case of a small test sample, a more restrictive adjustment leads to a non-significant outperformance of the SVM classifier due to a higher standard error and a decreasing difference of the two compared areas under the ROC. However, in case of a small training sample, the outperformance of the SVM classifier is statistically significant for all outlier adjustment levels.

Obviously, outlier adjustments have a larger impact on the classification performance of a LRM. It can be concluded that the specification of the separating hyperplane is less biased by outliers than the estimation of the LRM. This underlines the generalization advantages of SVM classifiers. Furthermore, the SVM classifier yields a higher classification accuracy for smaller training sample sizes.

The comparison of the Brier scores of the PD prediction, including its decompositions reliability RL and resolution RS, are displayed in Table 5 for alternative outlier adjustments of the training data and in Table 6 for alternative outlier adjustments of the calibration data.

As the training data is adjusted for outliers, the prediction of PD by the SVM classifier is much more accurate for all test partitions and outlier adjustment levels ε . The lower Brier score results from a higher resolution and a highly superior reliability. A more restrictive adjustment of the training data improves the resolution of the LRM, but deteriorates its reliability. In contrast, the reliability and resolution of the SVM classifier is hardly affected. The Brier score of the SVM classifier based PD prediction is nearly constant for all outlier adjustment levels and test partitions, while the goodness of the LRM-based PD forecasts is worsened by a more restrictive adjustment.

As shown in Table 6, the SVM classifier also allows a more accurate prediction of PD, when the calibration data is adjusted for outliers. The variance of the default frequency

Table 5 Goodness of the estimation of the probability of default (adjusted training data)

Partitioning training/testing	ϵ (%)	U	SVM			LRM		
			BS	RL	RS	BS	RL	RS
75%/25%	0	0.03453	0.03256	0.00001	0.00198	0.03375	0.00035	0.00114
	0.5	0.03453	0.03249	0.00002	0.00207	0.03393	0.00101	0.00161
	1	0.03453	0.03251	0.00002	0.00204	0.03425	0.00130	0.00159
	2	0.03453	0.03246	0.00002	0.00209	0.03433	0.00143	0.00164
83%/17%	0	0.03453	0.03239	0.00001	0.00215	0.03381	0.00050	0.00122
	0.5	0.03453	0.03238	0.00001	0.00217	0.03389	0.00093	0.00158
	1	0.03453	0.03251	0.00002	0.00205	0.03427	0.00130	0.00157
	2	0.03453	0.03244	0.00006	0.00215	0.03433	0.00141	0.00162
90%/10%	0	0.03453	0.03244	0.00001	0.00210	0.03366	0.00044	0.00131
	0.5	0.03453	0.03235	0.00003	0.00221	0.03392	0.00099	0.00160
	1	0.03453	0.03245	0.00004	0.00212	0.03421	0.00125	0.00157
	2	0.03453	0.03246	0.00003	0.00210	0.03431	0.00140	0.00162

Source: Authors calculation. The Brier score (BS) and its decompositions uncertainty (variance) of the default frequency (U) as well as the reliability (RL) and the resolution (RS) of the PD estimation are given for different partition settings and outlier adjustment of the training data at level ϵ

Table 6 Goodness of the estimation of the probability of default (adjusted calibration data)

Partitioning training/testing	ϵ (%)	U	SVM			LRM		
			BS	RL	RS	BS	RL	RS
75%/25%	0	0.03453	0.03256	0.00001	0.00198	0.03375	0.00035	0.00114
	0.5	0.03347	0.03157	0.00002	0.00192	0.03246	0.00015	0.00117
	1	0.03242	0.03044	0.00002	0.00200	0.03124	0.00008	0.00126
	2	0.03134	0.02925	0.00003	0.00211	0.03013	0.00005	0.00125
83%/17%	0	0.03453	0.03256	0.00001	0.00198	0.03375	0.00035	0.00114
	0.5	0.03347	0.03157	0.00002	0.00192	0.03246	0.00015	0.00117
	1	0.03242	0.03044	0.00002	0.00200	0.03124	0.00008	0.00126
	2	0.03134	0.02925	0.00003	0.00211	0.03013	0.00005	0.00125
90%/10%	0	0.03453	0.03244	0.00001	0.00210	0.03366	0.00044	0.00131
	0.5	0.03347	0.03148	0.00001	0.00201	0.03231	0.00020	0.00136
	1	0.03242	0.03034	0.00001	0.00209	0.03109	0.00010	0.00143
	2	0.03134	0.02919	0.00002	0.00218	0.02995	0.00006	0.00145

Source: Authors calculation. The Brier score (BS) and its decompositions uncertainty (variance) of the default frequency (U) as well as the reliability (RL) and the resolution (RS) of the PD estimation are given for different partition settings and outlier adjustment of the calibration data at level ϵ

decreases from 0.03453 to 0.03134 due to the exclusion of outliers from the calibration data. Hence, the lower Brier score mainly results from a sample-driven decrease in variance, but also from superior resolution and reliability of the PD prediction. A more restrictive adjustment of the calibration data improves the resolution and reliability of the LRM based PD forecasts. The reliability and resolution of the predicted PD based on the

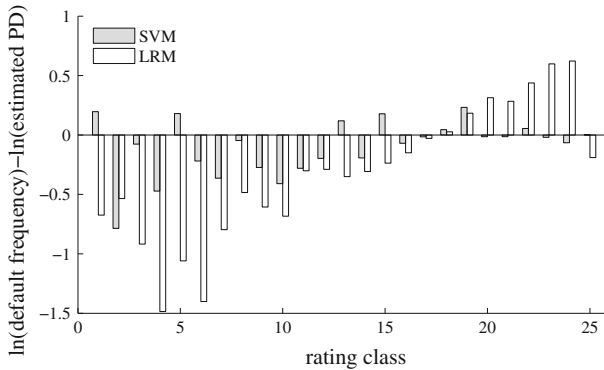


Fig. 2 Logarithmically standardized prediction error LRL_i of all rating classes $i = 1, \dots, 25$ without any outlier adjustments. 75% of the data sample is allocated for training and 25% for testing

SVM classifier are hardly affected. Hence, the absolute outperformance of the SVM classifier in terms of resolution and reliability decreases.

Concerning the PD prediction, it may be summarized that the PD prediction based on the SVM classifier is obviously less biased by outliers, as already observed for classification. An accurate PD prediction by the LRM requires an outlier adjustment of the calibration data. In contrast, the adjustment of the training data improves the classification performance of the LRM, but worsens the goodness of PD forecasting. Surprisingly, a variation of the partitioning of training and test data hardly affects the goodness of probability prediction. This holds true for the SVM classifier and LRM.

Finally Fig. 2 illustrates the logarithmically standardized prediction error LRL_i for all rating classes i . The PD forecasts based on the SVM classifier are much more reliable in most of the rating classes. The LRM leads to an overestimation of the probability of default in *good* rating classes and to an underestimation in *bad* rating classes. The plots of the estimated PDs and the observed default frequencies for a non-adjusted (a) and a 2% adjusted (b) calibration sample based on the LRM and the SVM classifier are illustrated in Figs. 3 and 4. The shape of the default frequency distribution and the curve of the estimated PD indicate that the rating score s provided by the SVM classifier leads to a better ranking order of default and non-default companies. The default frequencies were much

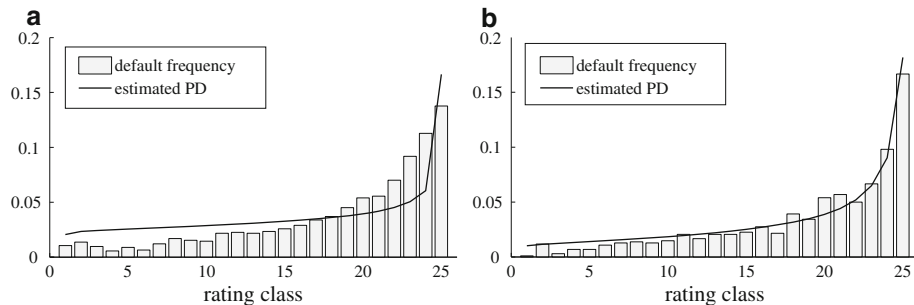


Fig. 3 Default frequency and estimated probability of default based on the LRM for an outlier adjustment at level $\epsilon = 0\%$ (a) and $\epsilon = 2\%$ (b). 75% of the data sample is allocated for training and 25% for testing

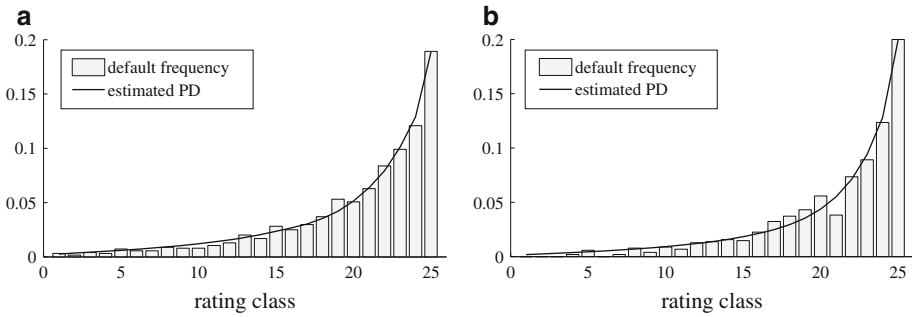


Fig. 4 Default frequency and estimated probability of default based on the SVM classifier for an outlier adjustment at level $\varepsilon = 0\%$ (a) and $\varepsilon = 2\%$ (b). 75% of the data sample is allocated for training and 25% for testing

better estimated by the PD forecasts based on the SVM classifier. Furthermore, the non-linear fitting of model (26) based on the rating score provided by the LRM is suboptimal, if the calibration data has a high variance and is not adjusted for outliers. In contrast, the SVM classifier allows a good fitting of model (26) for adjusted and non-adjusted calibration data.

5 Conclusion

This paper shows that SVM classifiers are powerful learning systems which are suitable for default classification and the estimation of probabilities of default (PD). Our empirical

Table 7 Results of the Wald test and the likelihood ratio test for the estimated LRM

Partitioning training/testing	ε (%)	Prob. value of the β -coefficient					LR-test
		CFROI	NDR	IC	ER	LT	
75%/25%	0	0.0933	0.4522	0.7236	0.4966	0.0000	0.0000
	0.5	0.0000	0.0268	0.3320	0.0000	0.0000	0.0000
	1	0.0000	0.0001	0.0404	0.0000	0.0000	0.0000
	2	0.0001	0.0010	0.0030	0.0000	0.0000	0.0000
83%/17%	0	0.0550	0.0398	0.6585	0.7857	0.0000	0.0000
	0.5	0.0000	0.0125	0.8118	0.0000	0.0000	0.0000
	1	0.0000	0.0003	0.0194	0.0000	0.0000	0.0000
	2	0.0000	0.0103	0.0015	0.0000	0.0000	0.0000
90%/10%	0	0.0047	0.7864	0.8145	0.2296	0.0000	0.0000
	0.5	0.0000	0.0306	0.4272	0.0000	0.0000	0.0000
	1	0.0000	0.0001	0.0542	0.0000	0.0000	0.0000
	2	0.0000	0.0103	0.0020	0.0000	0.0000	0.0000

Source: Authors calculation. The β -coefficients of the default indicators CFROI, NDR, IC, ER and LT were statistically tested by the Wald test. The significance of the indicator coefficients is given by the probabilistic values of the Wald test. Probabilistic values out of the 5% confidence level are tabulated in bold face. The goodness-of-fit of the proposed LRM is statistically tested by the likelihood ratio test. LR-Test is the probabilistic value of the likelihood ratio test. All statistics are given for alternative partition settings and outlier adjustment levels ε of the training data

results show that the theoretical advantages of SVM classifiers can be used to improve the accuracy of default classification and the reliability of PD prediction. The specification of an optimal separating hyperplane is obviously less sensitive to outliers than the maximum likelihood estimation of the LRM, and its classification performance and the PD prediction is superior, especially when sample size is limited. In contrast to the LRM, we found that the SVM classifier is more robust to a higher variance of the training and calibration data. Hence, SVM seem better suited for use with financial ratios, where variances are usually high and data availability often is a bottleneck. Although the parameterization and selection of the kernel function remains a very complex issue, we find that in sum, SVM classifiers are superior to LRM in default classification and default probability prediction.

A Statistics of the LRM

The likelihood ratio test of the LRM indicates a very good fit of the model. While the default indicator LT is highly significant for all data allocation settings, the other predictor variables are only statistically significant at the 5% level, if the training data is restrictively adjusted for outliers. Especially the default indicator IC is a critical variable. Because an exclusion of this indicator from the LRM worsens the PD forecasts in rating classes with low default frequencies, this predictor was not removed from the model.

The goodness-of-fit of the model is highly significant for all data allocation settings. The probabilistic values of the coefficient test (two-tailed Wald-Test) and of the likelihood ratio test of the LRM (LR-Test) are displayed in Table 7.

References

- Abe S (2005) Support vector machines for pattern classification. Springer, London
- Atiya AF (2001) Bankruptcy prediction for credit risk using neural networks: a survey and new results. *IEEE Trans Neural Netw* 12(4):929–935
- Baesens B, van Gestel T, Viaene S, Stepanova M, Suykens JAK, Vanthienen J (2003) Benchmarking state-of-the-art classification algorithms for credit scoring. *J Oper Res Soc* 54:627–635
- Bamber D (1975) The area above ordinal dominance graph and the area below the receiver operating characteristic graph. *J Math Psychol* 12:387–415
- Basel Committee on Banking Supervision (2006) International convergence of capital measurement and capital standards. Bank for International Settlements, Basel
- Boser BE, Guyon IM, Vapnik VN (1992) A training algorithm for optimal margin classifiers. In: Haussler D (ed) Proceedings of the 5th annual ACM workshop on computational learning theory. ACM Press, New York, pp 144–152
- Butera G, Faff R (2006) An integrated multi-model credit rating system for private firms. *Rev Quantitat Finance Account* 26:311–340
- Carling K, Jacobson T, Lind T J, Roszbach K (2007) Corporate credit risk modeling and the macroeconomy. *J Bank Finance* 31(3):845–868
- Chen LH, Chiou TW (1999) A fuzzy credit-rating approach for commercial loans: a taiwan case. *Omega* 27:407–419
- Chen S, Härdle W, Moro R (2006) Estimation of default probabilities with support vector machines. Discussion Paper 77, SFB 649 Humboldt University, Berlin
- Cortes C, Vapnik VN (1995) Support-vector networks. *Mach Learn* 20(3):273–297
- Cristianini N, Shawe-Taylor J (2006) An introduction to support vector machines. Cambridge University Press, Cambridge
- DeLong ER, DeLong DM, Clarke-Pearson DL (1988) Comparing the areas under two or more correlated receiver operating characteristic curves: a nonparametric approach. *Biometrics* 44:837–845
- Engelmann B, Hayden E, Tasche D (2003) Testing rating accuracy. *Risk* pp 82–86

- Evgeniou T, Pontil M, Poggio T (2000) Regularization networks and support vector machines. *Adv Comput Math* 13:1–50
- Hastie T, Tibshirani R (1990) *Generalized additive models*. Chapman and Hall, London
- Hastie T, Tibshirani R, Friedman J (2001) *The elements of statistical learning: data mining, inference, and prediction*. Springer, New York
- Hersbach H (2000) Decomposition of the continuous ranked probability score for ensemble prediction systems. *Am Meteorol Soc* 15:559–570
- Hosmer DW, Lemeshow S (2000) *Applied logistic regression*, 2nd edn. Wiley, New York
- Huang Z, Chen H, Hsu CJ, Chen WH, Wu S (2004) Credit rating analysis with support vector machines and neural networks: a market comparative study. *Decis Support Syst* 37:543–558
- Härdle W, Moro R, SchSfer D (2005) Predicting bankruptcy with support vector machines. In: Cizek P, Härdle W, Weron R (eds) *Statistical tools for finance and insurance*. Springer, Berlin, pp 225–248
- Härdle W, Lee YJ, SchSfer D, Yeh YR (2007) The default risk of firms examined with smooth support vector machines. Discussion Paper 757, DIW Berlin
- McLachlan GJ (2004) *Discriminant analysis and statistical pattern recognition*. Wiley, New York
- Ravi Kumar P, Ravi V (2007) Bankruptcy prediction in banks and firms via statistical and intelligent techniques—a review. *Eur J Oper Res* 180:1–28
- Schölkopf B, Smola AJ (2002) *Learning with kernels*. MIT Press, Cambridge
- Sun L (2007) A re-evaluation of auditors' opinions versus statistical models in bankruptcy prediction. *Rev Quantitat Finance Account* 28:55–78
- Suykens JA, Vandewalle J (1999) Least squares support vector machine classifiers. *Neural Process Lett* 9:293–300
- Suykens JA, van Gestel T, Brabanter JD, Moor BD, Vandewalle J (2002) *Least squares support vector machines*. World Scientific, Singapore
- Theodoridis S, Koutroumbas K (2006) *Pattern recognition*. Elsevier Academic Press, Amsterdam
- van Gestel T, Suykens JA, Baesens B, Viaene S, Vanthienen J, Dedene G, de Moor Joss Vandewalle B (2004) Benchmarking least squares support vector machine classifiers. *Mach Learn* 54:5–32
- Vapnik VN (1998) *Statistical learning theory*. Wiley, New York
- Vapnik VN (2000) *The nature of statistical learning theory*, 2nd edn. Springer, Berlin
- Varetto F (1998) Genetic algorithms applications in the analysis of insolvency risk. *J Bank Finance* 22:1421–1439
- Yobas MB, Crook JN, Ross P (2000) Credit scoring using neural and evolutionary techniques. *IMA J Manage Math* 11(2):111–125