Journal of Forecasting
J. Forecast. **28**, 512–534 (2009)
Published online 2 December 2008 in Wiley InterScience
(www.interscience.wiley.com) **DOI**: 10.1002/for.1109

Variable Selection and Oversampling in the Use of Smooth Support Vector Machines for Predicting the Default Risk of Companies

WOLFGANG HÄRDLE,[1] YUH-JYE LEE,[2]
DOROTHEA SCHÄFER[3]* AND YI-REN YEH[2]
[1] *CASE, Humboldt University, Berlin, Germany*
[2] *Department of Computer Science Information Engineering, National Taiwan University of Science and Technology, Taipei, Taiwan*
[3] *German Institute of Economic Research, Berlin, Germany*

ABSTRACT
In the era of Basel II a powerful tool for bankruptcy prognosis is vital for banks. The tool must be precise but also easily adaptable to the bank's objectives regarding the relation of false acceptances (Type I error) and false rejections (Type II error). We explore the suitability of smooth support vector machines (SSVM), and investigate how important factors such as the selection of appropriate accounting ratios (predictors), length of training period and structure of the training sample influence the precision of prediction. Moreover, we show that oversampling can be employed to control the trade-off between error types, and we compare SSVM with both logistic and discriminant analysis. Finally, we illustrate graphically how different models can be used jointly to support the decision-making process of loan officers.

KEY WORDS insolvency prognosis; support vector machines; statistical learning theory; non-parametric classification

INTRODUCTION

Default prediction is at the core of credit risk management and has therefore always attracted special attention. It has become even more important since the Basel Committee on Banking Supervision (Basel II) established borrowers' rating as the crucial criterion for minimum capital requirements of banks. The methods for generating rating figures have developed significantly over the last 10 years (Krahnen and Weber, 2001). The rationale behind the increased sophistication in predicting borrowers' default risk is the aim of banks to minimize their cost of capital and to mitigate their own bankruptcy risks.

* Correspondence to: Dorothea Schäfer, German Institute for Economic Research (DIW) Berlin, Mohrenstrasse 58, 10117 Berlin, Germany. E-mail: dschaefer@diw.de

In this paper we intend to contribute to the increasing sophistication by exploring the predicting power of smooth support vector machines (SSVM). SSVM are a variant of the conventional support vector machines (SVM). The working principle of SVM in general can be described very easily. Imagine a group of observations in distinct classes such as balance sheet data from solvent and insolvent companies. Assume that the observations are such that they cannot be separated by a linear function. Rather than fitting nonlinear curves to the data, SVM handle this problem by using a specific transformation function—the kernel function—that maps the data from the original space into a higher-dimensional space where a hyperplane can do the separation linearly. The constrained optimization calculus of SVM gives a unique optimal separating hyperplane and adjusts it in such a way that the elements of distinct classes possess the largest distance to the hyperplane. By re-transforming the separating hyperplane into the original space of variables, the typical nonlinear separating function emerges (Vapnik, 1995). The main difference between SSVM and SVM is the following: the SSVM technique formulates the problem as an unconstrained minimization problem. This formulation has mathematical properties such as strong convexity and desirable infinite differentiability.

Our aim is threefold when using SSVM. Firstly, we examine the power of the SSVM in predicting company defaults; secondly, we investigate how important factors that are exogenous to the model, such as selecting the appropriate set of accounting ratios, length of training period and structure of the training sample, influence the precision; and thirdly, we explore how oversampling and downsampling affect the trade-off between Type I and Type II errors. In addition, we illustrate graphically how loan officers can benefit from jointly considering the prediction results of different SSVM variants and different models.

There are basically three distinct approaches in predicting the risk of default: option theory-based approaches, parametric models and non-parametric methods. While the first class relies on the rule of no arbitrage, the latter two are based purely on statistic principles. The popular (Merton, 1974) model treats the company's equity as the underlying asset of a call option held by shareholders. In case of insolvency shareholders deny exercising. The probability of default is derived from an adapted Black–Scholes formula. Later, several authors (e.g., Longstaff and Schwartz, 1995; Mella-Barral and Perraudin, 1997; Leland and Toft, 1996; Zhou, 2001; to name only a few) proposed variations to ease the strict assumptions on the structure of the data imposed by the Merton model. These approaches are frequently denoted as structural models. However, the most challenging requirement is the knowledge of market values of debt and equity. This precondition is a severe obstacle to using the Merton model adequately as it is only satisfied in a minority of cases.

Parametric statistical models can be applied to any type of data, whether they are market based or book based. The first model introduced was discriminant analysis (DA) for univariate (Beaver, 1966) and multivariate models (Altman, 1968). After DA usage of the logit and probit approach for predicting default was proposed in Martin (1977) and Ohlson (1980). These approaches rely on the a priori assumed functional dependence between risk of default and predictor. DA requires a linear functional dependence, or a pre-shaped polynomial functional dependence in advanced versions. Logit and probit tools work with monotonic relationships between default event and predictors such as accounting ratios. However, such restrictions often fail to meet the reality of observed data. This fact makes it clear that there is a need for an approach that, in contrast to conventional methods, relaxes the requirements on data and/or lowers the dependence on heuristics. Semi-parametric models as in Hwang *et al.* (2007) are between conventional linear models and non-parametric approaches. Nonlinear classification methods such as support vector machines (SVM) or neural networks are even stronger candidates to meet these demands as they go beyond conventional

DOI: 10.1002/for

discrimination methods. Tam and Kiang (1992) and Altman *et al.* (1994) focus on neural networks. In contrast, we concentrate on SVM exclusively.

The SVM method is a relatively new technique and builds on the principles of statistical learning theory. It is easier to handle compared to neural networks. Furthermore, SVM have a wider scope of application as the class of SVM models includes neural networks (Schölkopf and Smola, 2002). The power of SVM technology becomes evident in a situation as depicted in Figure 1 where operating profit margin and equity ratio are used as explanatory variables. A separating function similar to a parabola (in black) appears in the two-dimensional space. The accompanying light-grey lines represent the margin boundaries whose shape and location determine the distance of elements from the separating function. In contrast, the logit approach and discriminant DA yield the (white) linear separating function (Härdle *et al.*, 2007a).

Selecting the best accounting ratios for executing the task of predicting is an important issue in practice but has not received appropriate attention in research. We address this issue of how important the chosen set of predictors is for the outcome. For this purpose we explore the prediction potential of SSVM within a two-step approach. First, we derive alternative sets of accounting ratios that are used as predictors. The benchmark set comes from Chen *et al.* (2006). A second set is defined by a 1-norm SVM, and the third set is based on the principle of adding only those variables that contain the most contrary information with respect to an initial set that is a priori chosen. We call the latter procedure the incremental forward selection of variables. As a result we are working with three variants of SSVM. In the second step, these variants are compared with respect to their prediction power. We also compare SSVM with two traditional methods: the logit model and linear discriminant analysis.

The analysis is built on 28 accounting ratios of 20,000 solvent and 1000 insolvent German companies. Our findings show that the different SSVM types have an overall good performance with the means of correct predictions ranging from 70% to 78%. The SSVM on the basis of incremental

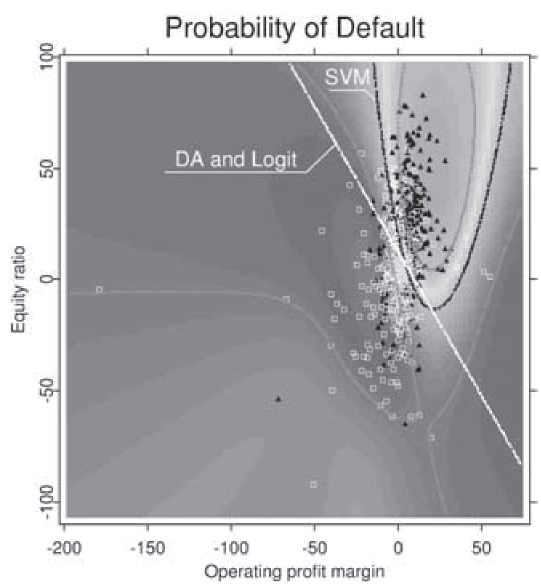

Figure 1. SVM-separating function (black) with margin in a two-dimensional space

DOI: 10.1002/for

forward selection clearly outperform the SSVM based on predictors selected by the 1-norm SVM. It is also found that oversampling influences the trade-off between Type I and Type II errors. Thus, oversampling can be used to make the relation of the two error types an issue of bank policy.

The rest of the paper is organized as follows. The following two sections describe the data, performance measures and SVM methodology. In the fourth section the variable selection technique and outcome are explained. The fifth section presents the experimental settings, estimation procedure and findings, and illustrates selected results. The sixth section concludes.

DATA AND MEASURES OF ACCURACY

In this study of the potential virtues of SVM in insolvency prognosis the CreditReform database is employed. The database consists of 20,000 financially and economically solvent and 1000 insolvent German companies observed once in the period from 1997 to 2002. Although the companies were randomly selected, accounting information dates most frequently in 2001 and 2002. Approximately 50% of the observations come from this period. The industry distribution of the insolvent companies is as follows: manufacturing 25.7%, wholesale and retail trade 20.1%, real estate 9.4%, construction 39.7% and others 5.1%. The latter includes businesses in agriculture, mining, electricity, gas and water supply, transport and communication, financial intermediation social service activities and hotels and restaurants. The 20,000 solvent companies belong to manufacturing (27.4%), wholesale and retail trade (24.8%), real estate (16.9%), construction (13.9%) and others (17.1%). There is only low coincidence between the industries represented in the insolvent and the solvent group of 'others'. The latter comprises many companies in industries such as publication administration and defense, education and health. Figure 2 shows the distribution of solvent and insolvent companies across industries. A set of balance sheet and income statement items describes each company. The ones we use for further analysis are described below:

- AD (amortization and depreciation)
- AP (accounts payable)
- AR (account receivable)

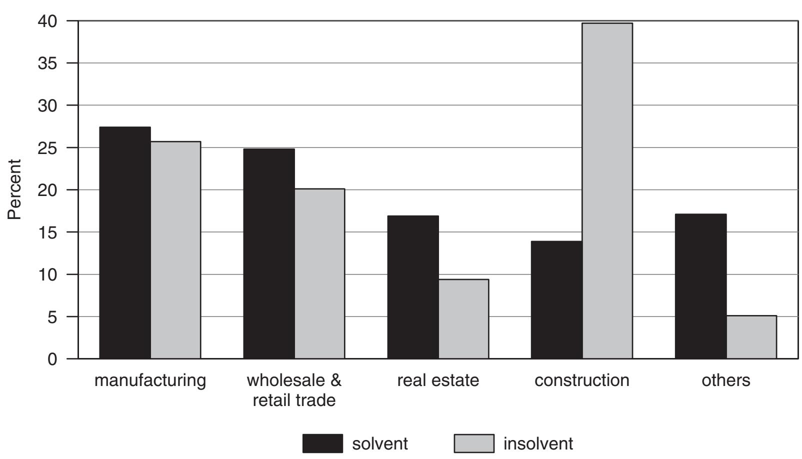

Figure 2. The distribution of solvent and insolvent companies across industries

- CA (current assets)
- CASH (cash and cash equivalents)
- CL (current liabilities)
- DEBT (debt)
- EBIT (earnings before interest and tax)
- EQUITY (equity)
- IDINV (growth of inventories)
- IDL (growth of liabilities)
- INTE (interest expense)
- INV (inventories)
- ITGA (intangible assets)
- LB (lands and buildings)
- NI (net income)
- OI (operating income)
- QA (quick assets)
- SALE (sales)
- TA (total assets)
- TL (total liabilities)
- WC (working capital (= CA − CL))

The companies appear in the database several times in different years; however, each year of balance sheet information is treated as a single observation. The data of the insolvent companies were collected 2 years prior to insolvency. The company sizes are measured by total assets. We construct 28 ratios to condense the balance sheet information (see Table I). However, before dealing with the CreditReform dataset, some companies whose behavior is very different from other ones are filtered out in order to make the dataset more compact. The data pre-processing procedure is described as follows:

1. We excluded companies whose total assets were not in the range of 10^5–10^7 EUR (remaining insolvent: 967; solvent: 15,834).
2. In order to compute the accounting ratios AP/SALE, OI/TA, TL/TA, CASH/TA, IDINV/INV, INV/SALE, EBIT/TA and NI/SALE, we have removed companies with zero denominators (remaining insolvent: 816; solvent 11,005).
3. We dropped outliers, that is, in the insolvent class companies with extreme values of financial indices have been removed (remaining insolvent: 811; solvent: 10,468).

After pre-processing, the dataset consists of 11,279 companies (811 insolvent and 10,468 solvent). In the following analysis, we focus on the revised dataset.

The performance of the SSVM is evaluated on the basis of three measures of accuracy: Type I error rate (%), Type II error rate (%) and total error rate (%). The Type I error is the ratio of the number of insolvent companies predicted as solvent ones to the number of insolvent companies. The Type II error is the ratio of the number of solvent companies predicted as insolvent ones to the number of solvent companies. Accordingly, the error-type rates (in percentage) are defined as follows

- Type I error rate = FN/(FN + TP) × 100 (%);
- Type II error rate = FP/(FP+ TN) × 100 (%);
- Total error rate = (FN + FP)/(TP + TN + FP + FN) × 100 (%);

DOI: 10.1002/for

Table I. Definitions of accounting ratios used in the analysis

Variable	Ratio	Indicator for
X1	NI/TA	Profitability
X2	NI/SALE	Profitability
X3	OI/TA	Profitability
X4	OI/SALE	Profitability
X5	EBIT/TA	Profitability
X6	(EBIT + AD)/TA	Profitability
X7	EBIT/SALE	Profitability
X8	EQUITY/TA	Leverage
X9	(EQUITY-ITGA)/	Leverage
	(TA-ITGA-CASH-LB)	Leverage
X10	CL/TA	Leverage
X11	(CL-CASH)/TA	Leverage
X12	TL/TA	Leverage
X13	DEBT/TA	Leverage
X14	EBIT/INTE	Leverage
X15	CASH/TA	Liquidity
X16	CASH/CL	Liquidity
X17	QA/CL	Liquidity
X18	CA/CL	Liquidity
X19	WC/TA	Liquidity
X20	CL/TL	Liquidity
X21	TA/SALE	Activity
X22	INV/SALE	Activity
X23	AR/SALE	Activity
X24	AP/SALE	Activity
X25	Log(TA)	Size
X26	IDINV/INV	Growth
X27	IDL/TL	Growth
X28	IDCASH/CASH	Growth

where

True positive (*TP*): Predict insolvent companies as insolvent ones
False positive (*FP*): Predict solvent companies as insolvent ones
True negative (*TN*): Predict solvent companies as solvent ones
False negative (*FN*): Predict insolvent companies as solvent ones

The following matrix explains the terms used in the definition of error rates:

		Predicted class	
		Positive	Negative
Actual	Positive	*True positive* (*TP*)	*False negative* (*FN*)
Class	Negative	*False positive* (*FP*)	*True negative* (*TN*)

SVM METHODOLOGY

In recent years, the so-called support vector machines (SVM), which have their roots in the theory of statistical learning (Burges, 1998; Christianini and Shawe-Taylor, 2000; Vapnik, 1995) have

DOI: 10.1002/for

become one of the most successful learning algorithms for classification as well as for regression (Drucker *et al.*, 1997; Mangasarian and Musicant, 2000; Smola and Schölkopf, 2004). Some features of SVM make them particularly attractive for predicting the default risk of companies. SVM are a non-parametric technique that learn the separating function from the data; they are based on a sound theoretical concept, do not require a particular distribution of the data, and deliver an optimal solution for the expected loss from misclassification. SVM estimate the separating hyperplane between defaulting and non-defaulting companies under the constraint of a maximal margin between the two classes (Vapnik, 1995; Schölkopf and Smola, 2002).

SVM can be formulated differently. However, in all variants either a constrained minimization problem or an unconstrained minimization problem is solved. The objective function in these optimization problems basically consists of two parts: a misclassification penalty part which stands for *model bias* and a regularization part which controls the *model variance*. We briefly introduce three different models: the smooth support vector machines (SSVM) (Lee and Mangasarian, 2001), the smooth support vector machines with reduced kernel technique (RSVM) and the 1-norm SVM. The SSVM will be used for classification and the 1-norm SVM will be employed for variable selection. The RSVM are applied for oversampling in order to mitigate the computational burden due to increasing the number of instances in the training sample.

Smooth support vector machines

The aim of the SVM technique is to find the separating hyperplane with the largest margin from the training data. This hyperplane is 'optimal' in the sense of statistical learning: it strikes a balance between overfitting and underfitting. Overfitting means that the classification boundary is too curved and therefore has less ability to classify unseen data correctly. Underfitting, on the other hand, gives a too simple classification boundary and leaves too many misclassified observations (Vapnik, 1995). We begin with linear support vector machines. Given a training dataset $S = \{(\mathbf{x}_1, y_1), \ldots, (\mathbf{x}_n, y_n)\} \subseteq \mathbb{R}^d \times \mathbb{R}$, where $\mathbf{x}_i \in \mathbb{R}^d$ is the input data and $y_i \in \{-1, 1\}$ is the corresponding class label, a conventional SVM separating hyperplane is generated by solving a convex optimization problem given as follows:

$$\begin{aligned} \min_{(w,b,\xi)\in\mathbb{R}^{d+1+n}} \; & C\sum_{i=1}^{b} \xi_i + \frac{1}{2}\|w\|_2^2 \\ \text{s.t.} \; & y_i(w^\top \mathbf{x}_i + b) + \xi_i \geq 1 \\ & \xi_i \geq 0, \quad \text{for } i = 1, 2, \ldots, n \end{aligned} \tag{1}$$

where C is a positive parameter controlling the trade-off between the training error (model bias) and the part of maximizing the margin (model variance) that is achieved by minimizing $\|w\|_2^2$. In contrast to the conventional SVM of (1), smooth support vector machines minimize the square of the slack vector ξ with weight $\frac{C}{2}$. In addition, the SSVM methodology appends $\frac{b^2}{2}$ to the term that is to be minimized. This expansion results in the following minimization problem:

$$\begin{aligned} \min_{(w,b,\xi)\in\mathbb{R}^{d+1+n}} \; & \frac{C}{2}\sum_{i=1}^{n} \xi_i^2 + \frac{1}{2}\left(\|w\|_2^2 + b^2\right) \\ \text{s.t.} \; & y_i(w^\top \mathbf{x}_i + b) + \xi_i \geq 1 \\ & \xi_i \geq 0, \quad \text{for } i = 1, 2, \ldots, n \end{aligned} \tag{2}$$

DOI: 10.1002/for

In a solution of (2), ξ is given by $\xi_i = \{1 - y_i(w^\top \mathbf{x}_i + b)\}_+$ for all i where the *plus* function x_+ is defined as $x_+ = \max\{0, x\}$. Thus, we can replace ξ_i in (2) by $\{1 - y_i(w^\top \mathbf{x}_i + b)\}_+$. This will convert the problem (2) into an unconstrained minimization problem as follows:

$$\min_{(w,b)\in\mathbb{R}^{d+1}} \frac{C}{2}\sum_{i=1}^{n}\{1-y_i(w^\top \mathbf{x}_i+b)\}_+^2+\frac{1}{2}\left(\|w\|_2^2+b^2\right) \tag{3}$$

This formulation reduces the number of variables from $d + 1 + n$ to $d + 1$. However, the objective function to be minimized is not twice differentiable, which precludes the use of a fast Newton method. In the SSVM, the plus function x_+ is approximated by a smooth *p-function*, $p(x,\alpha) = x + \frac{1}{\alpha}\log(1+e^{-\alpha x}), \alpha > 0$. Replacing the plus function with a very accurate smooth approximation p-function gives the smooth support vector machine formulation:

$$\min_{(w,b)\in\mathbb{R}^{d+1}} \frac{C}{2}\sum_{i=1}^{n} p(\{1-y_i(w^\top \mathbf{x}_i+b)\},\alpha)^2+\frac{1}{2}\left(\|w\|_2^2+b^2\right) \tag{4}$$

where $\alpha > 0$ is the smooth parameter. The objective function in problem (4) is strongly convex and infinitely differentiable. Hence, it has a unique solution and can be solved by using a fast Newton–Armijo algorithm. For the nonlinear case, this formulation can be extended to the nonlinear SVM by using the kernel trick as follows:

$$\min_{(u,b)\in\mathbb{R}^{n+1}} \frac{C}{2}\sum_{i=1}^{n} p\left(\left[1-y_i\left\{\sum_{j=1}^{n}u_j K(\mathbf{x}_i,\mathbf{x}_j)+b\right\}\right],\alpha\right)^2+\frac{1}{2}\left(\|u\|_2^2+b^2\right) \tag{5}$$

where $K(\mathbf{x}_i, \mathbf{x}_j)$ is a kernel function. This kernel function represents the inner product of $\phi(\mathbf{x}_i)$ and $\phi(\mathbf{x}_j)$, where ϕ is a certain mapping from input space $\mathbb{R}^d$ to a feature space $\mathcal{F}$. We do not need to know the mapping of ϕ explicitly. This is the so-called kernel trick. The nonlinear SSVM classifier can be expressed in matrix form as follows:

$$\sum_{u_j\neq 0} u_j K(A_j^\top, \mathbf{x}) + b = K(\mathbf{x}, A^\top)u + b \tag{6}$$

where $A = [\mathbf{x}_1^\top; \ldots; \mathbf{x}_n^\top]$ and $A_j = \mathbf{x}_j^\top$.

Reduced support vector machine

In large-scale problems, the full kernel matrix will be very large so it may not be appropriate to use the full kernel matrix when dealing with (5). In order to avoid facing such a big full kernel matrix, we brought in the reduced kernel technique (Lee and Huang, 2007). The key idea of the reduced kernel technique is to randomly select a portion of data and to generate a thin rectangular kernel matrix, then to use this much smaller rectangular kernel matrix to replace the full kernel matrix. In the process of replacing the full kernel matrix by a reduced kernel, we use the Nyström approximation (Smola and Schölkopf, 2000) for the full kernel matrix:

$$K(A, A^\top) \approx K(A, \tilde{A}^\top)K(\tilde{A}, \tilde{A}^\top)^{-1}K(\tilde{A}, A^\top) \tag{7}$$

DOI: 10.1002/for

where $K(A, A^\top) = K_{n\times n}$, $\tilde{A}_{\tilde{n}\times d}$ is a subset of A and $K(A, \tilde{A}) = \tilde{K}_{n\times\tilde{n}}$ is a reduced kernel. Thus, we have

$$K(A, A^\top)u \approx K(A, \tilde{A}^\top)K(\tilde{A}, \tilde{A}^\top)^{-1}K(\tilde{A}^\top, A)u = K(A, \tilde{A}^\top)\tilde{u} \tag{8}$$

where $\tilde{u} \in \mathbb{R}^{\tilde{n}}$ is an approximated solution of u via the reduced kernel technique. The reduced kernel method constructs a compressed model and cuts down the computational cost from $\mathcal{O}(n^3)$ to $\mathcal{O}(\tilde{n}^3)$. It has been shown that the solution of reduced kernel matrix approximates the solution of full kernel matrix well. The SSVM with the reduced kernel are called RSVM.

1-Norm support vector machine

The 1-norm support vector machine replaces the regularization term $\|w\|_2^2$ in (1) with the ℓ_1-norm of w. The ℓ_1-norm regularization term is also called the LASSO penalty (Tibshirani, 1996). It tends to shrink the coefficients w's towards zeros in particular for those coefficients corresponding to redundant noise features (Zhu *et al.*, 2003; Williams and Seeger, 2001). This nice feature will lead to a way of selecting the important ratios in our prediction model. The formulation of 1-norm SVM is described as follows:

$$\begin{aligned} \min_{(w,b,\xi)\in\mathbb{R}^{d+1+n}} \quad & C\sum_{i=1}^{n}\xi_i + \|w\|_1 \\ \text{s.t.} \quad & y_i(w^\top \mathbf{x}_i + b) + \xi_i \geq 1 \\ & \xi_i \geq 0, \quad \text{for } i = 1, 2, \ldots, n. \end{aligned} \tag{9}$$

The objective function of (9) is a piecewise linear convex function. We can reformulate it as the following linear programming problem:

$$\begin{aligned} \min_{(w,s,b,\xi)\in\mathbb{R}^{d+d+1+n}} \quad & C\sum_{i=1}^{n}\xi_i + \sum_{j=1}^{d} s_j \\ \text{s.t.} \quad & y_i(w^\top \mathbf{x}_i + b) + \xi_i \geq 1 \\ & -s_j \leq w_j \leq s_j, \quad \text{for } j = 1, 2, \ldots, d, \\ & \xi_i \geq 0, \quad \text{for } i = 1, 2, \ldots, n \end{aligned} \tag{10}$$

where s_j is the upper bound of the absolute value of w_j. In the optimal solution of (10) the sum of s_j is equal to $\|w\|_1$.

The 1-norm SVM can generate a very sparse solution w and lead to a parsimonious model. In a linear SVM classifier, solution sparsity means that the separating function $f(\mathbf{x}) = w^\top\mathbf{x} + b$ depends on very few input attributes. This characteristic can significantly suppress the number of nonzero coefficient w's, especially when there are many redundant noise features (Fung and Mangasarian, 2004; Zhu *et al.*, 2003). Therefore the 1-norm SVM can be a very promising tool for the variable selection tasks. We will use it to choose the important financial indices for our bankruptcy prognosis model.

SELECTION OF ACCOUNTING RATIOS

In principle any possible combination of accounting ratios could be used as explanatory variables in a bankruptcy prognosis model. Therefore, appropriate performance measures are needed to gear the process of variable selection towards picking the ratios with the highest separating power. In

DOI: 10.1002/for

Chen *et al.* (2006) accuracy ratio (AR) and conditional information entropy ratio (CIER) determine the selection procedure's outcome. It turned out that the ratio 'accounts payable divided by sales', X24 (AP/SALE), has the best performance values for a univariate SVM model. The second selected variable was the one combined with X24 that had the best performance in a bivariate SVM model. This is the analogue of forward selection in linear regression modeling. Typically, improvement declines if new variables are added consecutively. In Chen *et al.* (2006) the performance indicators started to decrease after the model included eight variables. The described selection procedure is quite lengthy, since there are at least 216 accounting ratio combinations to be considered. We will not employ the procedure here but use the chosen set of eight variables as the benchmark set V1. Table II presents V1 in the first column.

We propose two different approaches for variable selection that will simplify the selection procedure. The first one is based on 1-norm SVM introduced above. The SVM were applied to the period from 1997 to 1999. We selected the variables according to the size of the absolute values of the coefficients w from the solution of the 1-norm SVM. Table II displays the eight selected variables as V2. We obtain eight variables out of 28. Note that five variables, X2, X3, X5, X15 and X24, are also in the benchmark set V1.

The second variable selection scheme is incremental forward variable selection. The intuition behind this scheme is that a new variable will be added into the already selected set, if it brings in the most extra information. We measure the extra information for an accounting ratio using the distance between this new ratio vector and the space spanned by the current selected ratio subset. This distance can be computed by solving a least-squares problem (Lee *et al.*, 2008). The ratio with the farthest distance will be added into the selected accounting ratio set. We repeat this procedure until a certain stopping criterion is satisfied. The accounting ratio X24 (AP/SALE) is used as the initial selected accounting ratio. Then we follow the procedure seven times to select seven more extra accounting ratios. The variable set generated is called V3. We will use these three variable sets, V1, V2 and V3, for further data analysis in the next section. The symbol $^+$ denotes the variables that are common to all sets: X2, X3, X5 and X24.

Table II. Selected variables

Variable	Definition	V1	V2	V3
X2$^+$	NI/SALE	x	x	x
X3$^+$	OI/TA	x	x	x
X4	OI/SALE			x
X5$^+$	EBIT/TA	x	x	x
X6	(EBIT + AD)/TA		x	
X7	EBIT/SALE			x
X8	EQUITY/TA		x	
X12	TL/TA	x		
X13	DEBT/TA			x
X15	CASH/TA	x	x	
X21	TA/SALE			x
X22	INV/SALE	x		
X23	AR/SALE		x	
X24$^+$	AP/SALE	x	x	x
X26	IDINV/INV	x		

DOI: 10.1002/for

EXPERIMENTAL SETTING AND RESULTS

In this section we present our experimental setting and results. We compare the performance of three sets of accounting ratios, V1, V2 and V3, in our SSVM-based insolvency prognosis model. The performance is measured by Type I error rate, Type II error rate and total error rate. Fortunately, in reality, there is only a small number of insolvent companies compared to the number of solvent companies. Due to the small share in a sample that reflects reality, a simple classification such as naive Bayesian or a decision tree tends to classify every company as solvent. Such a classification would imply accepting all companies' loan applications and would thus lead to a very high Type I error rate while the total error rate and the Type II error rate are very small. Such models are useless in practice.

Our cleaned dataset consists of around 10% of insolvent companies. Thus, the sample is fairly unbalanced although the share of insolvent companies is higher than in reality. In order to deal with this problem, insolvency prognosis models usually start off with more balanced training and testing samples than reality can provide. For example, Härdle *et al.* (2007b) employ a downsampling strategy and work with balanced (50%/50%) samples. The chosen bootstrap procedure repeatedly randomly selects a fixed number of insolvent companies from the training set and adds the same number of randomly selected solvent companies. However, in this paper we adopt an oversampling strategy, to balance the size between the solvent and the insolvent companies, and refer to the downsampling procedure primarily for reasons of reference.

Oversampling duplicates the number of insolvent companies a certain number of times. In this experiment, we duplicate in each scenario the number of insolvent companies as many times as necessary to reach a balanced sample. Note that in our oversampling scheme every solvent and insolvent company's information is utilized. This increases the computational burden due to increasing the number of training instances. We employ the reduced kernel technique introduced above to mediate this problem.

All classifiers we need in these experiments are reduced SSVM with the Gaussian kernel, which is defined as

$$K(\mathbf{x}, \mathbf{z}) = e^{-\gamma\|\mathbf{x}-\mathbf{z}\|_2^2}$$

where γ is the width parameter. In nonlinear SSVM, we need to determine two parameters: the penalty term C and γ. The 2D grid search will consume a lot of time. In order to cut down the search time, we adopt the uniform design model selection method (Huang *et al.*, 2007) to search an appropriate pair of parameters.

Performance of SSVM

We conduct the experiments in a scenario in which we always train the SSVM bankruptcy prognosis model from the data at hand and then use the trained SSVM to predict the following year's cases. This strategy simulates the real task of prediction which binds the analyst to use past data for forecasting future outcomes. The experimental setting is described in Table III. The number of periods which enter the training set changes from 1 year (S1) to 5 years (S5).

In Tables IV and V we report the results for the oversampling and downsampling strategy respectively. Mean and standard deviation of Type I, Type II and total error rates (misclassification rates) are shown. We perform these experiments for the three variable sets, V1 to V3, and compare the oversampling and downsampling scheme in each experiment. All experiments are repeated 30 times

DOI: 10.1002/for

Table III. The scenario of our experiments

Scenario	Observation period of training set	Observation period of testing set
S1	1997	1998
S2	1997–1998	1999
S3	1997–1999	2000
S4	1997–2000	2001
S5	1997–2001	2002

Table IV. Results of oversampling for three variable sets (RSVM)

Set of accounting ratios	Scenario	Type I error rate		Type II error rate		Total error rate	
		Mean	SD	Mean	SD	Mean	SD
V1	S1	33.16	0.55	26.15	0.13	26.75	0.12
	S2	31.58	0.01	29.10	0.07	29.35	0.07
	S3	28.11	0.73	26.73	0.16	26.83	0.16
	S4	30.14	0.62	25.66	0.17	25.93	0.15
	S5	24.24	0.56	23.44	0.13	23.48	0.13
V2	S1	29.28	0.92	27.20	0.24	27.38	0.23
	S2	28.20	0.29	30.18	0.18	29.98	0.16
	S3	27.41	0.61	29.67	0.19	29.50	0.17
	S4	28.12	0.74	28.32	0.19	28.31	0.15
	S5	23.91	0.62	24.99	0.10	24.94	0.10
V3	S1	29.28	0.83	25.11	0.25	25.46	0.21
	S2	31.27	0.62	29.79	0.34	29.94	0.35
	S3	30.91	0.13	27.21	0.19	27.48	0.18
	S4	32.00	0.54	25.19	0.17	25.61	0.14
	S5	26.98	0.42	22.90	0.11	23.08	0.11

Table V. Results of downsampling for three variable sets (SSVM with Gaussian kernel)

Set of accounting ratios	Scenario	Type I error rate		Type II error rate		Total error rate	
		Mean	SD	Mean	SD	Mean	SD
V1	S1	32.20	3.12	28.98	1.70	29.26	1.46
	S2	29.74	2.29	28.77	1.97	28.87	1.57
	S3	30.46	1.88	26.23	1.33	26.54	1.17
	S4	31.55	1.52	23.89	0.97	24.37	0.87
	S5	28.81	1.53	23.09	0.73	23.34	0.69
V2	S1	29.94	2.91	28.07	2.15	28.23	1.79
	S2	28.77	2.58	29.80	1.89	29.70	1.52
	S3	29.88	1.88	27.19	1.32	27.39	1.19
	S4	29.06	1.68	26.26	1.00	26.43	0.86
	S5	26.92	1.94	25.30	1.17	25.37	1.06
V3	S1	30.87	3.25	26.61	2.45	26.98	2.11
	S2	33.31	2.16	28.60	2.01	29.08	1.65
	S3	31.82	1.52	26.41	1.45	26.80	1.31
	S4	35.0	2.13	24.29	0.77	24.96	0.68
	S5	30.66	1.60	21.92	0.96	22.30	0.92

DOI: 10.1002/for

because of the randomness in the experiments. The randomness is very obvious in the downsampling scheme (see Table V). Each time we only choose negative instances with the same size of the whole positive instances. The observed randomness in our oversampling scheme (Table IV) is due to applying the reduced kernel technique to solving the problem. We use the training set in the downsampling scheme as the reduced set. That is, we use all the insolvent instances and the equal number of solvent instances as our reduced set in generating the reduced kernel. Then we duplicate the insolvent part of the kernel matrix to balance the size of insolvent and solvent companies.

Both tables reveal that different variable selection schemes produce dissimilar results with respect to both precision and deviation of predicting. The oversampling scheme shows better results in the Type I error rate but has slightly bigger total error rates. It is also obvious that in almost all models a longer training period works in favor of accuracy of prediction. Clearly, the oversampling schemes have much smaller standard deviations in the Type I error rate, Type II error rate, and total error rate than the downsampling one. According to this observation, we conclude that the oversampling scheme will generate a more robust model than the downsampling scheme.

Figure 3 illustrates the development (learning curve) of the Type I error rate and total error rate with regard to variable set V3 for both oversampling and downsampling. The bullets on the lines

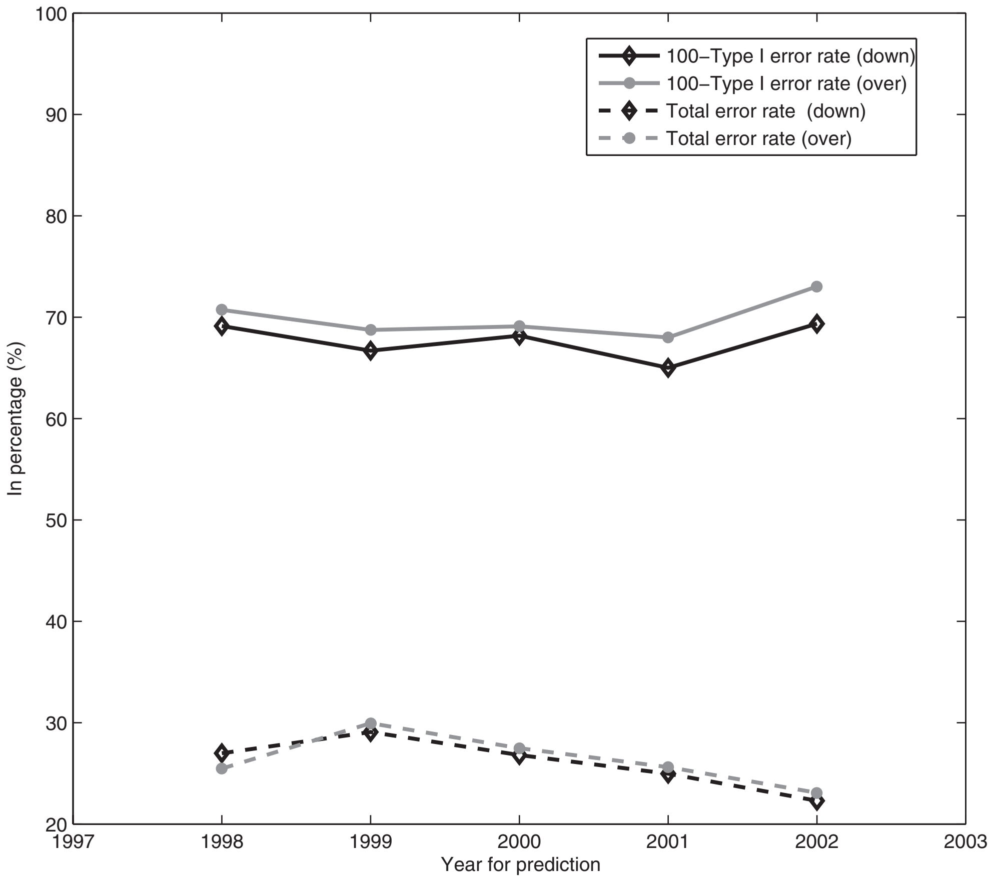

Figure 3. Learning curve for variables set V3

DOI: 10.1002/for

mark the different training scenarios. For example, the first bullets from the left represent S1 (training set from 1997, testing set from 1998), the second bullets illustrate S2 (training set from 1997 to 1998, testing set from 1999) etc. For the purpose of better visibility, the Type I error rate is only indirectly displayed as 100 – Type I error rate. The upper solid line in gray represents the oversampling scheme and the black solid line the downsampling one. Note that the performance in terms of the Type I error rate is worse the higher the distance between the upper end of the diagram and the solid lines. The learning curve over the time frame the training sample covers shows an upward tendency between S1 and S5 for the number 100 – Type 1 error rate. However, the curves are non-monotonic. There is a disturbance for the forecast of year 1999 that is based on training samples that cover 1997 to 1998, and also one for the forecast of year 2001 based on training samples covering 1997 to 2000. Both disturbances may have been caused by the reform of the German insolvency code that came into force in 1999. The most important objective of the reform was to allow for more company restructuring and less liquidation than before. This reform considerably changed the behavior of German companies towards declaring insolvency, and thus most likely the nature of balance sheets that are associated with insolvent companies.

The disturbances are less visible with respect to the overall performance. The dashed lines near the lower edge of the diagram box show total error rates, gray for the oversampling and black for the downsampling scheme. There is a clear tendency towards a lower total error rate from S2 to S5 for both schemes. The downsampling line is slightly below the oversampling one, representing a slightly better performance in terms of the mean of the total error rate. However, this result has to be seen in the light of the trade-off between magnitude and stability of results, as oversampling yields much more stable results. The standard deviations for V3 are only a small portion of the numbers generated by the downsampling procedure across all training scenarios (Tables IV and V).

Table VI presents the comparison between the sets by focusing on the total error rate. It indicates by an asterisk whether the differences in means are significant at the 10% level via *t*-test and, in addition, gives the set which is superior in the dual comparison. Variable set V2 is nearly absent in Table VI. Thus V2 is clearly outperformed by both sets V1 and V3. There is no clear distinction between V1 and V3 except for Scenario S5. Given the long training period V3 is superior in both the downsampling and oversampling scenarios and generates the lowest total error rate in absolute terms.

In order to investigate the effect of the oversampling versus the downsampling scheme we follow the setting as above, but we use the V3 variable set. For each training–test pair, we carry out oversampling for positive instances from 6 to 15 times. We show the trend and effect in Figure 4. It is

Table VI. Statistical significance in differences in means (10% level) between the three variable sets: total error

Sets	S1	S2	S3	S4	S5
Oversampling					
V1 vs. V2	V1*	V1*	V1*	V1*	V1*
V1 vs. V3	V3*	V1*	V1*	V3*	V3*
V2 vs. V3	V3*		V3*	V3*	V3*
Downsampling					
V1 vs. V2	V2*	V1*	V1*	V1*	V1*
V1 vs. V3	V3*			V1*	V3*
V2 vs. V3	V3*		V3*	V3*	V3*

DOI: 10.1002/for

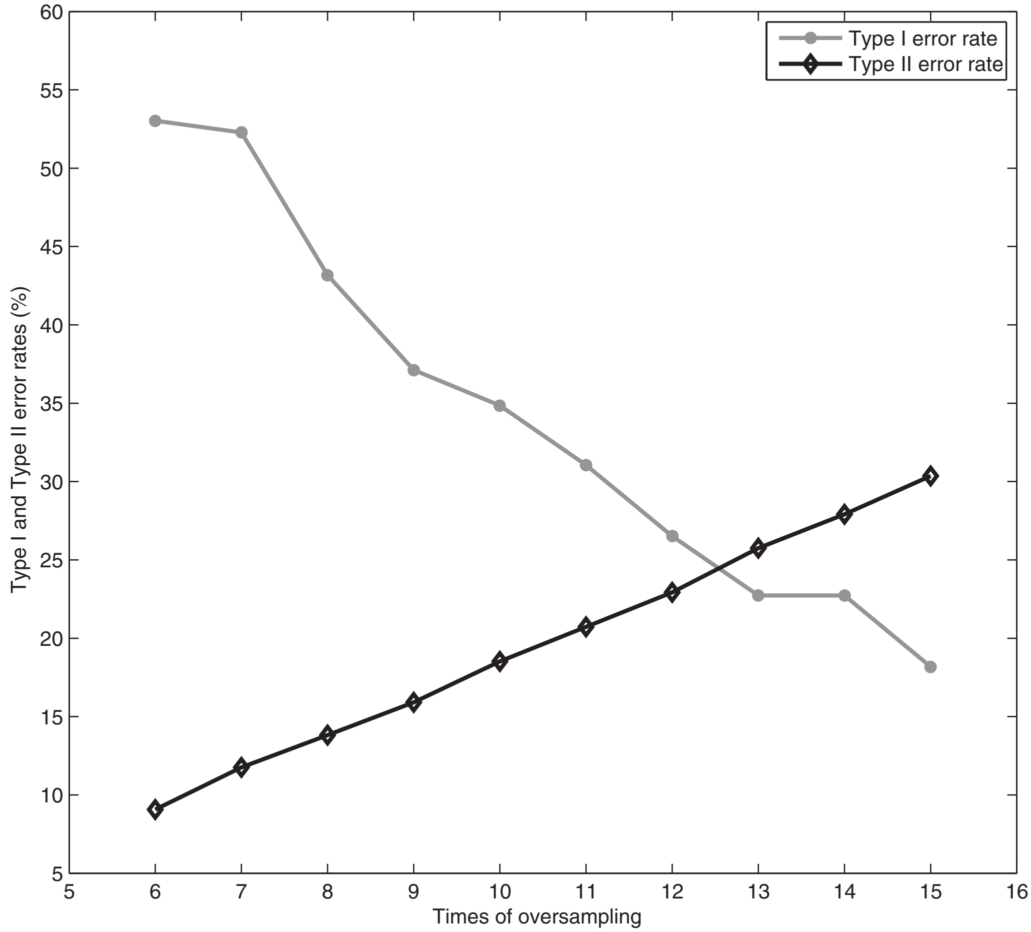

Figure 4. The effect of oversampling on Type I and Type II error rates for scenario S5 and variables set V3

easy to see that the Type I (II) error rate decreases (increases) as the oversampling times increase. This feature implies that the machine would have a tendency of classifying all companies as solvent if the training sample had realistic shares of insolvent and solvent companies. Such behavior would produce a Type I error rate of 100%. The more balanced the sample is, the higher the penalty for classifying insolvent companies as solvent. This fact is illustrated in Figure 4 by the decreasing curve with respect to the number of duplications of insolvent companies.

Often banks favor a strategy that allows them to minimize the Type II errors for a given number of Type I errors. The impact of oversampling on the trade-off between the two types of errors—shown in Figure 4—implies that the number of oversampling times is a strategic variable in training the machine. This number can be determined by the bank's aim regarding the relation of Type I and Type II errors.

DOI: 10.1002/for

Comparison with logit and linear discriminant analysis

The examination of SSVM is incomplete without comparing it to highly used traditional methods such as the logistic model (LM) and linear discriminant analysis (DA). Therefore, we replicate the research design of the previous section with both traditional models. In addition, we test whether the difference in means in the total error rate is statistically significant. The comparison of means with regard to the total error rate is presented in Tables VII and VIII for the oversampling and downsampling strategy respectively. Table IX summarizes the comparison of the approaches and displays the statistical significance of their mean differences. Asterisks indicate the out-performance

Table VII. Comparison of the total error rate (%) as generated by SSVM with LM and DA: oversampling for three variable sets

Set of accounting ratios	Scenario	SSVM	LM	DA
		Mean	Mean	Mean
V1	S1	26.75	26.50	25.60
	S2	29.35	28.96	27.22
	S3	26.83	28.94	27.42
	S4	25.93	26.20	25.55
	S5	23.48	26.95	28.23
V2	S1	27.38	26.80	26.20
	S2	29.98	28.63	28.70
	S3	29.50	29.52	29.46
	S4	28.31	28.43	28.08
	S5	24.94	29.22	31.42
V3	S1	25.46	25.07	23.65
	S2	29.94	28.29	27.02
	S3	27.48	27.89	25.84
	S4	25.61	26.60	24.85
	S5	23.08	25.32	26.15

Table VIII. Comparison of the total error rate (%) as generated by SSVM with LM and DA: downsampling for three variable sets

Set of accounting ratios	Scenario	SSVM	LM	DA
		Mean	Mean	Mean
V1	S1	29.26	26.86	27.34
	S2	28.87	28.62	28.26
	S3	26.54	27.54	28.22
	S4	24.37	24.80	25.47
	S5	23.34	24.81	25.86
V2	S1	28.23	27.28	28.62
	S2	29.70	29.29	29.65
	S3	27.39	28.56	29.58
	S4	26.43	26.41	27.96
	S5	25.37	26.52	29.69
V3	S1	26.98	26.03	25.47
	S2	29.08	28.04	27.22
	S3	26.80	26.60	26.51
	S4	24.96	25.25	25.44
	S5	22.30	23.96	24.31

DOI: 10.1002/for

Table IX. Statistical significance in differences of means (10% level) between SSVM and LM and SSVM and DA, respectively, for the sets V1 to V3: total error rate

V1	S1	S2	S3	S4	S5
Oversampling					
SSVM vs. LM			*	*	*
SSVM vs. DA			*		*
Downsampling					
SSVM vs. LM			*	*	*
SSVM vs. DA			*	*	*
V2	S1	S2	S3	S4	S5
Oversampling					
SSVM vs. LM				*	*
SSVM vs. DA					*
Downsampling					
SSVM vs. LM			*		*
SSVM vs. DA			*	*	*
V3	S1	S2	S3	S4	S5
Oversampling					
SSVM vs. LM			*	*	*
SSVM vs. DA					*
Downsampling					
SSVM vs. LM					*
SSVM vs. DA				*	*

of the logistic model or discriminant analysis by SSVMs at the 10% level via *t*-test. It is obvious that the SSVM technique yields the better results, the longer the period is from which the training observations are taken. In fact, the results show that the SSVM works significantly better than LM and DA in most cases in S3 to S5, with the clearest advantage for testing sets S4 and S5, where the accounting information of the predicted companies dates most frequently in 2001 and 2002.

We also investigate the effect of oversampling on LM and DA. We follow the same setting in the previous section, doing oversampling for positive instances from 6 to 15 times. Unlike the SSVM-based insolvency prognosis model, the DA approach is insensitive in both Type I and Type II error rates to the replication of positive instances. The result for DA is illustrated in Figure 5. The LM approach has very similar results to the SSVM model. We will not show the result here.

More data visualization

Each SSVM model has its own output value. We use this output to construct 2D coordinate systems. Figure 6 shows an example for scenario S5 where the scores of the $SSVM_{V3}$ model ($SSVM_{V1}$ model) are represented by the horizontal (vertical) line. A positive (negative) value indicates predicted insolvency (solvency). We then map all insolvent companies in the testing set onto the coordinate systems. There are 132 insolvent companies and 2866 solvent companies in this testing set. We also randomly choose the same amount of solvent companies from the testing set.

The plus points in the lower left quadrant and the circle points in the upper right quadrant show the number of Type I errors and Type II errors, respectively, in both models. Plus points in the upper right quadrant and circle points in the lower left quadrant reflect those companies that are predicted

DOI: 10.1002/for

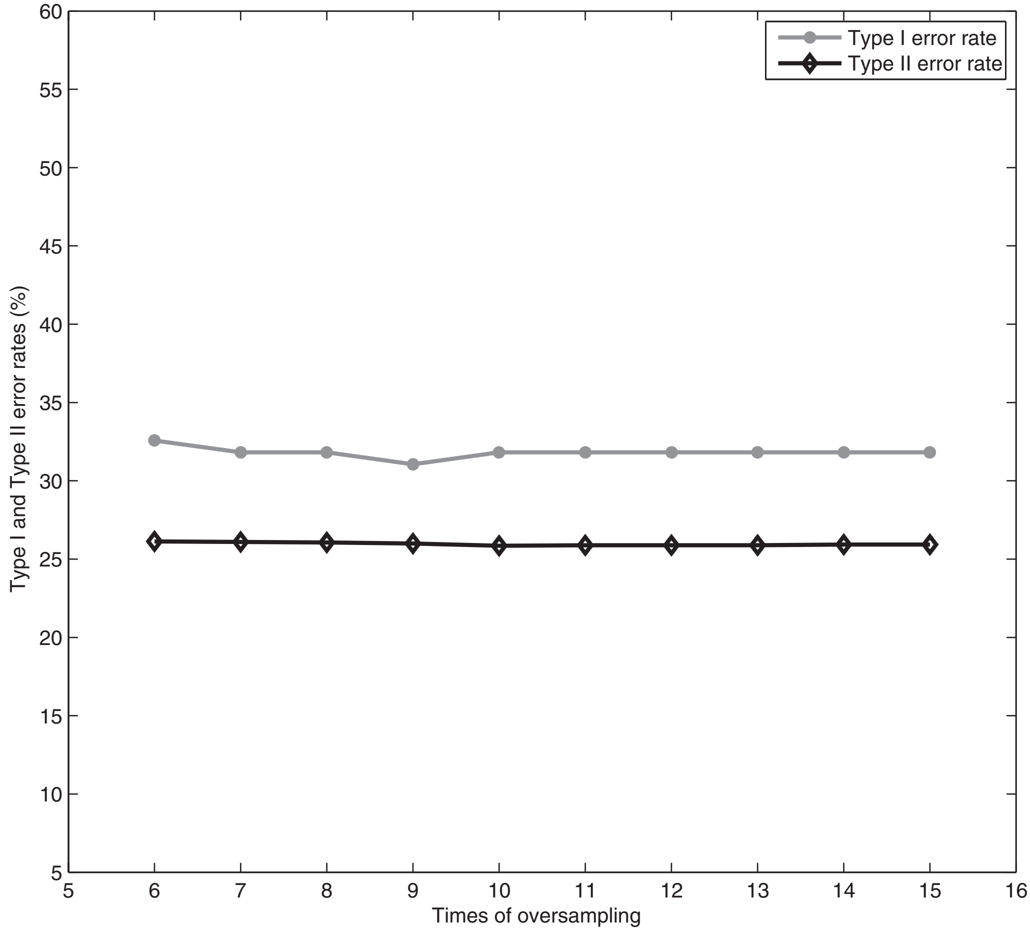

Figure 5. The effect of oversampling on Type I and Type II error rates for scenario S5 and variables set V3 in DA

correctly by both models. Circles and plus points in the lower right quadrant (upper left quadrant) represent conflicting prognoses. We also report the number of insolvent companies and the number of solvent companies in each quadrant of Figure 6. The two different insolvency prognosis models based on V1 and V3, respectively, can be considered as alternative experts. The two forecasts for each instance in the testing set is plotted in the diagram. The proposed visualization scheme could be used to support loan officers in their final decision on accepting or rejecting a client's application. Furthermore, this data visualization scheme can also be applied to two different learning algorithms, such as $SSVM_{V3}$ vs. LM_{V3} and $SSVM_{V3}$ vs. DA_{V3}. We show these data visualization plots in Figures 7 and 8. If the loan application has been classified as solvent or insolvent by alternative machines, it is most likely that the prognosis meets reality (the plus points in the upper right quadrant and the circle points in the lower left quadrant). Opposing forecasts, however, should be taken as a hint to evaluate the particular company more thoroughly, for example by employing an expert team, or even by using a third model.

DOI: 10.1002/for

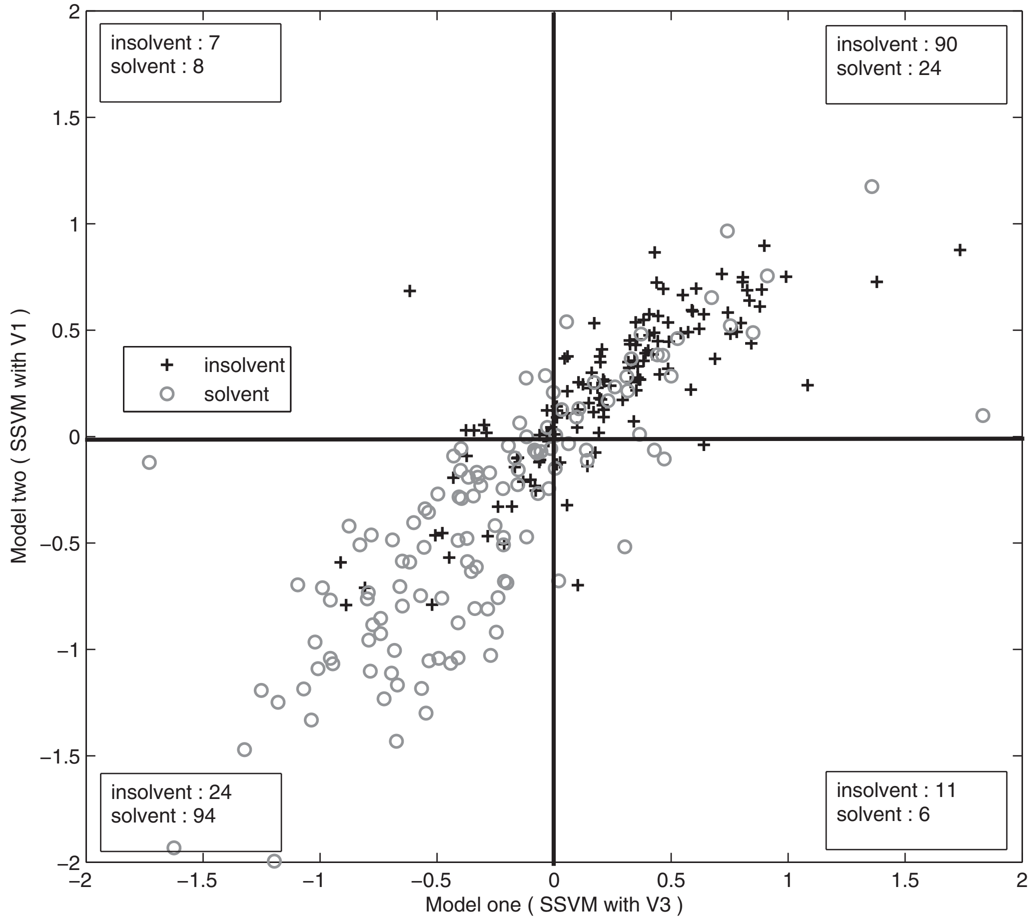

Figure 6. Data visualization via model one (generated by SSVM with V3) and model two (generated by SSVM with V1) in scenario S5

CONCLUSION

In this paper we apply different variants of support vector machines to a unique dataset of German solvent and insolvent companies. We use a priori a given set of predictors as a benchmark, and suggest two further variable selection procedures; the first procedure uses the 1-norm SVM and the second, incremental way consecutively selects the variable that is the farthest one from the column space of the current variable set. Given the three SSVM based on distinct variable sets, the relative performance of the types of smooth support vector machines is tested. The performance is measured by error rates. The two sets of variables newly created here lead to a dissimilar performance of SSVM. The selection of variables by the 1-norm SVM clearly underperforms compared to the incremental selection scheme. This difference in accuracy hints at the need for further research with respect to the variable selection. The training period makes a clear difference, though. Results improve considerably if more years of observation are used in training the machine. The SSVM

DOI: 10.1002/for

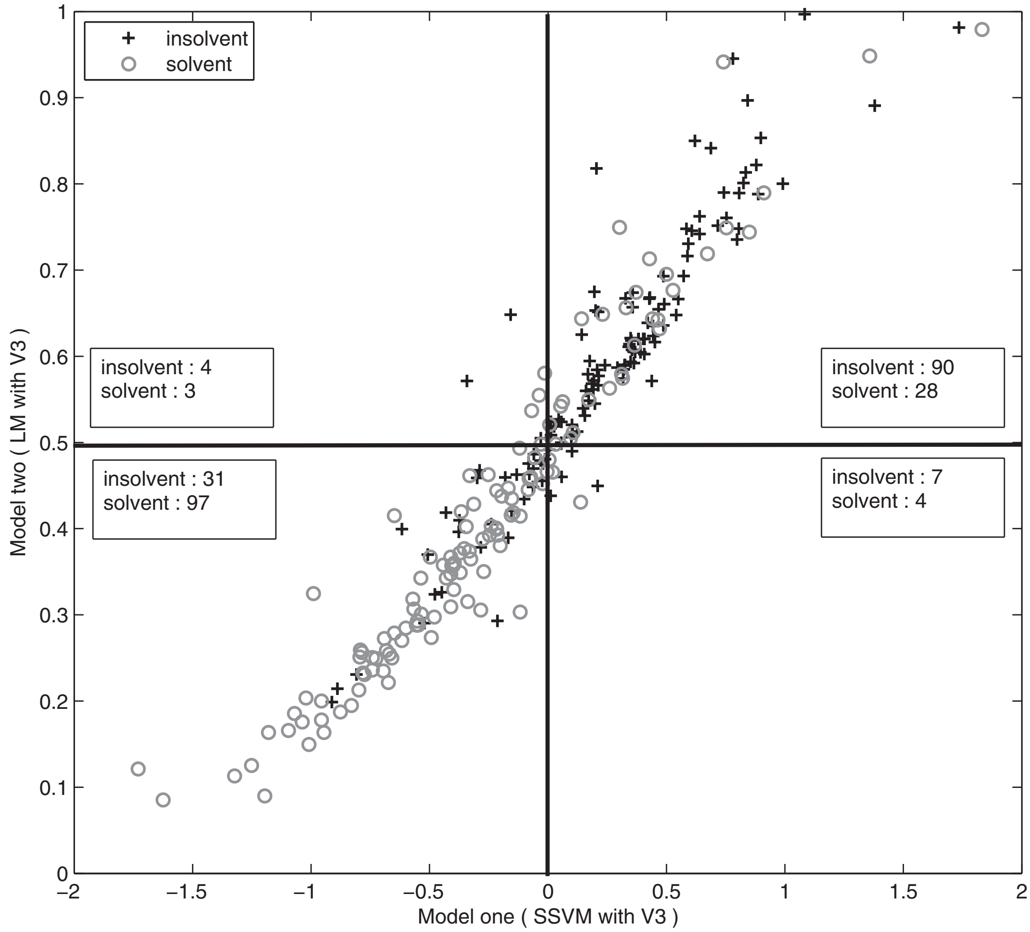

Figure 7. Data visualization via model one (generated by SSVM with V3) and model two (generated by LM with V3) in scenario S5

model benefits more from longer training periods than traditional methods do. As a consequence the logit model and discriminant analysis are both outperformed by the SSVM in long-term training scenarios. Moreover, the oversampling scheme works very well in dealing with unbalanced datasets. It provides flexibility to control the trade-off between Type I and Type II errors, and is therefore a strategic instrument in a bank's hand. The results generated are very stable in terms of small deviations of Type I, Type II and total error rates.

Finally, we want to stress that SSVM should be considered not as a substitute for traditional methods but rather as a complement which, when employed side by side with either the logit model or discriminant analysis, can generate new information that helps practitioners select those companies that are difficult to predict and, therefore, need more attention and further treatment.

DOI: 10.1002/for

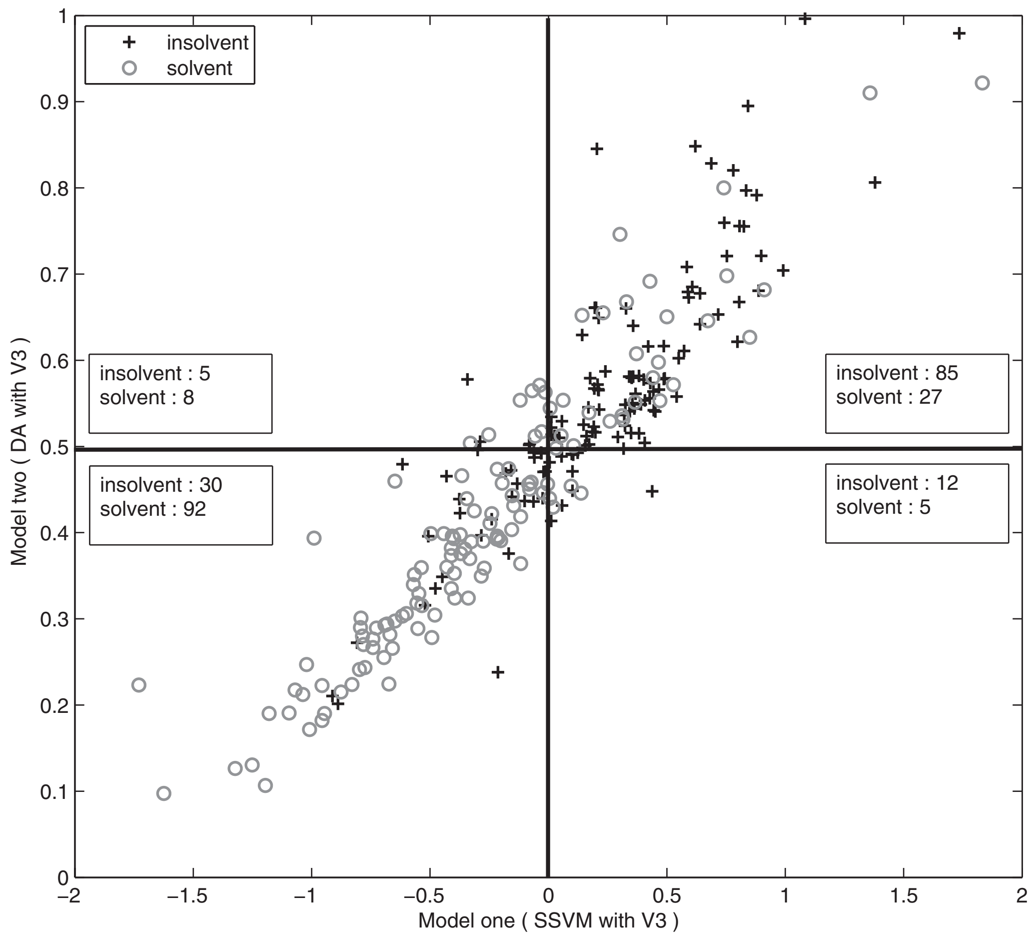

Figure 8, Data visualization via model one (generated by SSVM with V3) and model two (generated by DA with V3) in scenario S5

ACKNOWLEDGEMENTS

This research was supported by the 'Stiftung Geld und Währung' and by the Deutsche Forschungsgemeinschaft through the SFB 649 'Economic Risk'.

REFERENCES

Altman E. 1968. Financial ratios, discriminant analysis and the prediction of corporate bankruptcy. *Journal of Finance* **23**(4): 589–609.

Altman E, Marco G, Varetto F. 1994. Corporate distress diagnosis: comparisons using linear discriminant analysis and neural networks (the italian experience). *Journal of Banking and Finance* **18**: 505–529.

Beaver W. 1966. Financial ratios as predictors of failures: empirical research in accounting: selected studies. *Journal of Accounting Research* **4**: 71–111.

Burges CJC. 1998. A tutorial on support vector machines for pattern recognition. *Data Mining and Knowledge Discovery* **2**(2): 121–167.

Chen S, Härdle W, Moro RA. 2006. Estimation of default probabilities with support vector machines. *SFB 649 Discussion Paper 2006-077*.

Cristianini N, Shawe-Taylor J. 2000. *An Introduction to Support Vector Machines*. Cambridge University Press: Cambridge, UK.

Drucker H, Burges CJC, Kaufman L, Smola A, Vapnik V. 1997. Support vector regression machines. In *Advances in Neural Information Processing Systems 9*, Mozer MC, Jordan MI, Petsche T (eds). MIT Press: Cambridge, MA; 155–161.

Fung G, Mangasarian OL. 2004. A feature selection Newton method for support vector machine classification. *Computational Optimization and Applications* **28**(2): 185–202.

Härdle W, Moro R, Schäfer D. 2007a. Graphical data representation in bankruptcy analysis based on support vector machines. In *Handbook of Data Visualization*, Chen C, Härdle W, Unwin A (eds). Springer: Heidelberg; 853–872.

Härdle W, Moro RA, Schäfer D. 2007b. Estimating probabilities of default with support vector machines. *SFB 649 Discussion Paper 2007-035*.

Huang CM, Lee YJ, Lin DKJ, Huang SY. 2007. Model selection for support vector machines via uniform design. *Computational Statistics and Data Analysis* **52**: 335–346. Special Issue on Machine Learning and Robust Data Mining (to appear).

Hwang RC, Cheng KF, Lee JC. 2007. A semiparametric method for predicting bankruptcy. *Journal of Forecasting* **26**(5): 317–342.

Krahnen JP, Weber M. 2001. Generally accepted rating principles: a primer. *Journal of Banking and Finance* **25**(1): 3–23.

Lee YJ, Huang SY. 2007. Reduced support vector machines: a statistical theory. *IEEE Transactions on Neural Networks* **18**: 1–13.

Lee YJ, Mangasarian OL. 2001. SSVM: a smooth support vector machine. *Computational Optimization and Applications* **20**: 5–22.

Lee YJ, Chang CC, Chao CH. 2008. Incremental forward feature selection with application to microarray gene expression. *Journal of Biopharmaceutical Statistics* **18**(5): 824–840.

Leland H, Toft K. 1996. Optimal capital structure, endogenous bankruptcy, and the term structure of credit spreads. *Journal of Finance* **51**: 987–1019.

Longstaff FA, Schwartz ES. 1995. A simple approach to valuating risky fixed and floating rate debt. *Journal of Finance* **50**: 789–819.

Mangasarian OL, Musicant DR. 2000. Robust linear and support vector regression. *IEEE Transactions on Pattern Analysis and Machine Intelligence* **22**(9): 950–955.

Martin D. 1977. Early warning of bank failure: a logit regression approach. *Journal of Banking and Finance* **1**: 249–276.

Mella-Barral P, Perraudin W. 1997. Strategic debt service. *Journal of Finance* **52**: 531–556.

Merton R. 1974. On the pricing of corporate debt: the risk structure of interest rates. *Journal of Finance* **29**(2): 449–470.

Ohlson J. 1980. Financial ratios and the probabilistic prediction of bankruptcy. *Journal of Accounting Research* **18**(1): 109–131.

Schölkopf B, Smola AJ. 2002. *Learning with Kernels*. MIT Press: Cambridge, MA.

Smola A, Schölkopf B. 2000. Sparse greedy matrix approximation for machine learning. In *Proceedings of the 17th International Conference on Machine Learning*, San Francisco, CA.

Smola A, Schölkopf B. 2004. A tutorial on support vector regression. *Statistics and Computing* **14**: 199–222.

Tam K, Kiang M. 1992. Managerial application of neural networks: the case of bank failure prediction. *Management Science* **38**(7): 926–947.

Tibshirani R. 1996. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society* **58**(1): 267–288.

Vapnik VN. 1995. *The Nature of Statistical Learning Theory*. Springer: New York.

Williams CKI, Seeger M. 2001. Using the Nyström method to speed up kernel machines. *Advances in Neural Information Processing Systems* **13**: 682–688.

DOI: 10.1002/for

Zhou C. 2001. The term structure of credit spreads with jump risk. *Journal of Banking and Finance* **25**: 2015–2040.

Zhu J, Rosset S, Hastie T, Tibshirani R. 2003. 1-Norm support vector machines. In *Advances in Neural Information Processing Systems* **16**: 49–56.

Authors' biographies:

Wolfgang Härdle did in 1982 his Dr. rer. nat. in Mathematics at Universität Heidelberg and in 1988 his Habilitation at Universität Bonn. He is currently chair professor of statistics at the Dept. of Economics and Business Administration, Humboldt-Universität zu Berlin. He is also director of CASE—Center for Applied Statistics & Economics and of the Collaborative Research Center 'Economic Risk'. His research focuses on dimension reduction techniques, computational statistics and quantitative finance. He has published 34 books and more than 200 papers in top statistical, econometrics and finance journals. He is one of the 'Highly cited Scientist' according to the Institute of Scientific Information.

Yuh-Jye Lee received his Master degree in Applied Mathematics from the National Tsing Hua University, Taiwan in 1992 and PhD degree in computer sciences from the University of Wisconsin-Madison in 2001. In 2002, Dr. Lee joined the Computer Science and Information Engineering Department, National Taiwan University of Science and Technology. He is an associate professor now. His research interests are in machine learning, data mining, optimization, information security and operations research. He developed new algorithms for large data mining problems such as classification problem, clustering, feature selection and dimension reduction. These algorithms have been used in intrusion detection systems (IDS), face detection, micro array gene expression analysis and breast cancer diagnosis and prognosis.

Dorothea Schäfer did in 1992 her Dr. rer. pol. in Economics and in the year 2000 her Habilitation at Freie Universität Berlin. She is currently coordinator of the research group Financial Markets and Financial Institutions and senior researcher at the German Institute for Economic Research (DIW) Berlin which she joined in 2002. She is managing editor of the Quarterly Journal of Economic Research (Vierteljahreshefte zur Wirtschaftsforschung) and adjunct lecturer at Freie Universität Berlin. Her research focuses on insolvency risk, financial management of firms and banks, and on behavioural finance.

Yi-Ren Yeh received the M.S. degree from the Department of Computer Science and Information Engineering, National Taiwan University of Science and Technology, Taipei, Taiwan, R.O.C., in 2006. He is currently working toward the PhD degree in the Department of Computer Science and Information Engineering, National Taiwan University of Science and Technology. His research interests include machine learning, data mining, and information security.

Authors' addresses:

Wolfgang Härdle, Center for Applied Statistics and Economics, Humboldt-Universität zu Berlin, Spandauer Straße 1, 10178 Berlin, Germany.

Yuh-Jye Lee and **Yi-Ren Yeh**, Department of Computer Science Information Engineering, National Taiwan University of Science and Technology, Taipei 106, Taiwan.

Dorothea Schäfer, German Institute for Economic Research (DIW) Berlin, Mohrenstrasse 58, 10117 Berlin, Germany.

DOI: 10.1002/for