ECONSTOR
Make Your Publications Visible.

A Service of
ZBW
Leibniz-Informationszentrum Wirtschaft
Leibniz Information Centre for Economics

Härdle, Wolfgang Karl; Moro, Rouslan A.; Hoffmann, Linda

Working Paper

Learning machines supporting bankruptcy prediction

SFB 649 Discussion Paper, No. 2010,032

Provided in Cooperation with:
Collaborative Research Center 649: Economic Risk, Humboldt University Berlin

Suggested Citation: Härdle, Wolfgang Karl; Moro, Rouslan A.; Hoffmann, Linda (2010) : Learning machines supporting bankruptcy prediction, SFB 649 Discussion Paper, No. 2010,032, Humboldt University of Berlin, Collaborative Research Center 649 - Economic Risk, Berlin

This Version is available at:
https://hdl.handle.net/10419/39293

Standard-Nutzungsbedingungen:

Die Dokumente auf EconStor dürfen zu eigenen wissenschaftlichen Zwecken und zum Privatgebrauch gespeichert und kopiert werden.

Sie dürfen die Dokumente nicht für öffentliche oder kommerzielle Zwecke vervielfältigen, öffentlich ausstellen, öffentlich zugänglich machen, vertreiben oder anderweitig nutzen.

Sofern die Verfasser die Dokumente unter Open-Content-Lizenzen (insbesondere CC-Lizenzen) zur Verfügung gestellt haben sollten, gelten abweichend von diesen Nutzungsbedingungen die in der dort genannten Lizenz gewährten Nutzungsrechte.

Terms of use:

Documents in EconStor may be saved and copied for your personal and scholarly purposes.

You are not to copy documents for public or commercial purposes, to exhibit the documents publicly, to make them publicly available on the internet, or to distribute or otherwise use the documents in public.

If the documents have been made available under an Open Content Licence (especially Creative Commons Licences), you may exercise further usage rights as specified in the indicated licence.

WWW.ECONSTOR.EU

Mitglied der Leibniz-Gemeinschaft

SFB 649 Discussion Paper 2010-032

Learning Machines Supporting Bankruptcy Prediction

Wolfgang Karl Härdle*
Rouslan Moro**
Linda Hoffmann*

* Humboldt-Universität zu Berlin, Germany
** Brunel University, UK

This research was supported by the Deutsche Forschungsgemeinschaft through the SFB 649 "Economic Risk".

http://sfb649.wiwi.hu-berlin.de
ISSN 1860-5664

SFB 649, Humboldt-Universität zu Berlin
Spandauer Straße 1, D-10178 Berlin

SFB 649 ECONOMIC RISK BERLIN

Learning Machines Supporting Bankruptcy Prediction *

Wolfgang Karl Härdle†, Rouslan Moro‡and Linda Hoffmann §

June 7, 2010

Abstract

In many economic applications it is desirable to make future predictions about the financial status of a company. The focus of predictions is mainly if a company will default or not. A support vector machine (SVM) is one learning method which uses historical data to establish a classification rule called a score or an SVM. Companies with scores above zero belong to one group and the rest to another group.

Estimation of the probability of default (PD) values can be calculated from the scores provided by an SVM. The transformation used in this paper is a combination of weighting ranks and of smoothing the results using the PAV algorithm. The conversion is then monotone.

This discussion paper is based on the Creditreform database from 1997 to 2002. The indicator variables were converted to financial ratios; it transpired out that eight of the 25 were useful for the training of the SVM. The results showed that those ratios belong to activity, profitability, liquidity and leverage.

Finally, we conclude that SVMs are capable of extracting the necessary information from financial balance sheets and then to predict the future solvency or insolvent of a company. Banks in particular will benefit from these results by allowing them to be more aware of their risk when lending money.

*The financial support from the Deutsche Forschungsgemeinschaft via SFB 649 'Ökonomisches Risiko", Humboldt-Universität zu Berlin is gratefully acknowledged.

†Humboldt-Universität zu Berlin, C.A.S.E. - Center for Applied Statistics and Economics, Spandauer Str. 1, 10178 Berlin, Germany, and National Central University, Department of Finance, No. 300, Jhongda. Rd., Jhongli City, Taoyuan County 32001, Taiwan (R.O.C.). Email: haerdle@wiwi.hu-berlin.de

‡Doctor at the Department of Economics and Finance, School of Social Sciences, Brunel University. Email: Russ.Moro@brunel.ac.uk

§Humboldt-Universität zu Berlin, C.A.S.E. - Center for Applied Statistics and Economics, Spandauer Str. 1, 10178 Berlin, Germany. Email: linda.hoffmann@cms.hu-berlin.de

JEL classification: C14; G33; C45

Keywords: Support Vector Machine, Bankruptcy, Default Probabilities Prediction, Profitability

1 Introduction

This work presents one of the more recent and efficient learning systems – support vector machines (SVMs). SVMs are mainly used to classify various specialized categories such as object recognition (Schölkopf 1997)), optical character recognition (Vapnik (1995)), electric load prediction (Eunite (2001)), management fraud detection (Rätsch and Müller (2004)),and early medical diagnostics. It is also used to predict the solvency or insolvency of companies or banks, which is the focus of this work. In other words, SVMs are capable of extracting useful information from financial data and then label companies by giving them score values. Furthermore, probability of default (PD) values for companies can be calculated from those score values. The method is explained later.

In the past, discriminant analysis (DA) and logit models were used for classification, especially in the financial world. The logit model is in a way a generalized DA model because it does not assume multivariate normality and equal covariance matrices. Over time, researchers, bankers and others found out that those two models were inefficient because they cannot classify satisfactorily if the data is non-linear separable. Consequently, the rate of prediction of new companies was low and demand for more accurate default estimations was developed, thus w ork with artificial neural nets (ANN), decision trees and SVMs started. The literature (S. Chen (accepted in 2009)) has shown that SVMs produce better classification results than parametric methods. Additionally, SVMs have single solutions characterized by the global minimum of the optimized function and they do not rely heavily on heuristics. SVM are attractive estimators and aree therefore worth studying.

2 Bankruptcy analysis

Lending money is an act based on trust in the debtors ability to repay a loan. From where does the lender get this trust? Banks and other institutions rely heavily on statistical tools that try to predict the financial situation of borrowers. Since those tools are only estimates of reality, lenders risk losing what they have invested. Therefore, improving predictions of bankruptcies allow for better lending decisions. The task for statisticians is to bolster existing methods and to develop new ones.

The problem of default and credit risks is not new and the idea of using financial ratios to analyze companies is more than a century old. Ramser & Foster (1931), Fitzpatrick (1932) and Winakor & Smith (1935) were one of the first scientists to apply financial ratios for bankruptcy predictions. The systematic application of statistics to bankruptcy analysis began with the work of Beaver (1966) and Altman (1968). They introduced the univariate and multivariate discriminant analysis (DA). In 1968, Altman presented a formula for predicting bankruptcy known as the linear Z-score model. This formula was widely popular for calculating defaults and even today it is used due to its simplicity. A drawback of the Z-score model is the assumption of equal normal distributions for both failing and successful companies with the same covariance matrix. In reality, the distributions may not be normal, therefore, financial institutions required a more sophisticated method.

The centre of research shifted towards the logit and probit models. In 1977, Martin introduced 'Early warning of bank failure' and a few years later Ohlson (1980) published 'Financial ratios and the probabilistic prediction of bankruptcy.' Wiginton (1980), Zavgren (1983) and Zmijewski (1984) continued working on logit and probit models. During that time, other statisticians proposed different methods such as the gambler's ruin model (Wilcox 1971), option pricing theory (Merton 1974), recursive partitioning (Frydman, Altman & Kao 1985), neural networks (Tam & Kiang 1992) and rough sets (Dimitras, Slowinski, Susmaga & Zopounidis 1999). Glennon and Nigro (2005) suggested a hazard or survival analysis.

From a geometrical point of view, SVMs classify solvent and insolvent companies into two groups by putting a margin of separation between them. For the best classification, the margin needs to be maximized and the error of misclassification minimized. The easiest way to classify occurs when the data is linearly separable. However, this is not always the case. Sometimes two groups cannot be separated linearly in the dimension they exist but can be in a higher dimensional space. The kernel technique (Hastie, Tibshirani, and Friedman (2001)) allows us to map the data into a higher dimensional feature space. For that reason a SVM is a more powerful tool than classical a DA, logit or probit. (The two later are only linear classifiers.)

The purpose of an SVM is to classify new data x after training a classification function f. This f needs to be a good approximation of y when x is observed minimizing the expected risk

$$R(f) = \int |f(x) - y| \, dP(x, y). \tag{1}$$

y are the labels of x and $|f(x) - y|$ is known as the loss function. f is an element of the set of measurable functions $\mathcal{F}$. To avoid overfitting $\mathcal{F}$ is restricted to a smaller amount of functions. This approach is called emprirical risk minimization principle (ERM). In practice the distribution $P(x, y)$ is unknown. Therefore, $R(f)$ cannot be calculated but may be approximated.

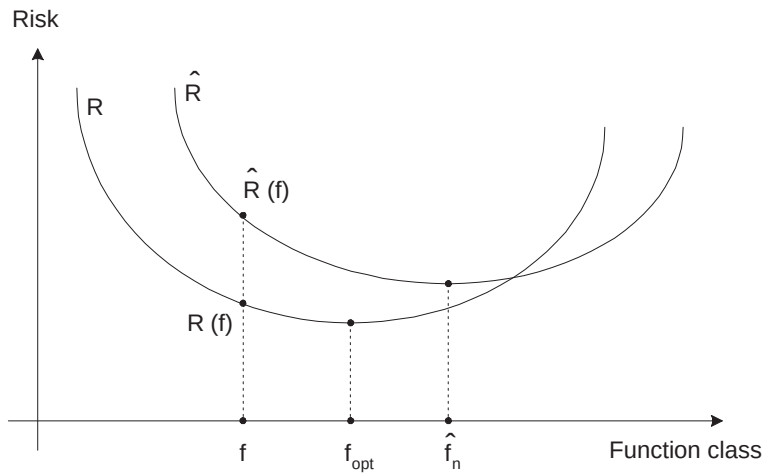

Figure 1: The minimum values f_{opt} and $\hat{f}_n$ of the expected (R) and empirical ($\hat{R}$) risk functions generally do not coincide.

Because of the unknown distribution $P(x, y)$, the empirical risk needs to be introduced:

$$\hat{R}(f) = \frac{1}{n} \sum_{i=1}^{n} |f(x_i) - y_i| . \tag{2}$$

Our loss function is the average value of misclassifications over the training set. Many loss functions exist such as the least square $\left(L_{\mathrm{LS}(y,t)} \stackrel{\mathrm{def}}{=} (1 - yt)^2\right)$, hinge $\left(L_{\mathrm{hinge}}(y, t) \stackrel{\mathrm{def}}{=} \max\{0, 1 - yt\}\right)$ and logistic $\left(L_{\mathrm{logic}(y,t)} \stackrel{\mathrm{def}}{=} \log\{1 + \exp(-yt)\}^2\right)$. These are used according to the problem posed or user preference.

The minimization of the expected and empirical risk:

$$f_{opt} = \arg\min_{f \in \mathcal{F}} R(f) , \tag{3}$$

$$\hat{f}_n = \arg\min_{f \in \mathcal{F}} \hat{R}(f) , \tag{4}$$

do not necessary coincide (Figure 1). Depending on the size of $\mathcal{F}$, f_{opt} and $\hat{f}_n$ will become arbitrarily close as n increases.

From this we conclude that minimizing the expected risk directly is not possible due to the unknown distribution $P(x, y)$. According to statistical learning theory (Vapnik 1995), it is possible to estimate the Vapnik-Chervonenkis (VC) bound by putting an upper bound on $R(f)$. With probability $1 - \eta$:

$$R(f) \leq \hat{R}(f) + \sqrt{\frac{h\left(log\frac{2n}{h} + 1\right) - log(\eta/4)}{n}}, \tag{5}$$

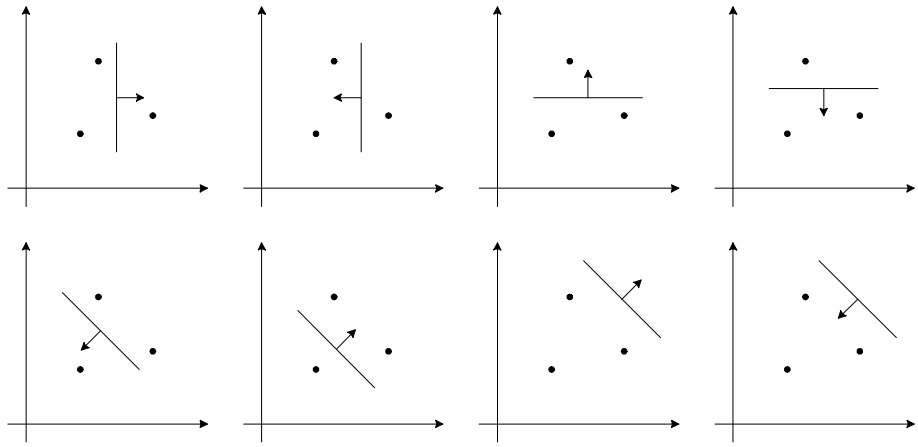

Figure 2: Eight possible ways of shattering 3 points on the plane with a linear indicator function.

where

$$h \leq min\left\{\left(r^2 \left\|w\right\|\right) + 1, n + 1\right\} \text{ (Vapnik 95).} \tag{6}$$

r is the radius of the smallest sphere containing data and $\frac{2}{\|w\|}$ is the width of the margin. h is the VC dimension.

Next we review the VC dimension and its relation to the SVM. If for some $f \in \mathcal{F}$, the objects $\left\{x_i \in \mathbb{R}^d, i = 1, ..., h\right\}$ can be shattered in all 2^h possible ways and no set $\left\{x_j \in \mathbb{R}^d, j = 1, ..., q\right\}$ exists with $q > n$, then h is called the VC dimension of $\mathcal{F}$ in a d-dimensional space. For instance, x_1, x_2 and $x_3 \in \mathbb{R}^2$ can be shattered linearly in $8 = 2^3$ ways. If we add a fourth point x_4, we cannot shatter them into $2^4 = 16$ different ways. Hence, the VC-dimension h is 3 in $\mathbb{R}^2$ as shown in Figure 2.

The expression for the VC bound (5) is a regularized functional where the VC dimension h is a parameter controlling the complexity of the classifier function. We could find a function that makes no training error but its performance on new data would be low. Therefore, it is important to control the complexity. This means we have a trade-off between the number of classification errors on the training set and the complexity of the classifier function.

The second goal is to maximize the margin separating the two groups as illustrated in Figure 3. The separating function generated by a linear SVM is

$$x^\top w + b = 0, \tag{7}$$

where w is of dimension $d \times 1$ called the weight or the slope. b is a scalar representing

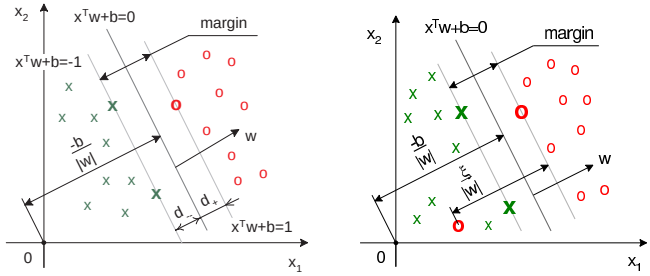

Figure 3: The separating hyperplane $x^\top w + b = 0$ and the margin in a linearly separable (left) and non-separable (right) case. Crosses denote solvent companies, zeros are the insolvent ones. The hyperplanes bounding the margin zone equidistant from the separating hyperplane are represented as $x^\top w + b = 1$ and $x^\top w + b = -1$. The misclassification penalty in the non-separable case is proportional to the distance $\xi / \|w\|$.

the location parameter. x_i is a $d \times 1$ vector of the characteristics of company i, e.g. financial ratios described in 4. There are d different characteristics.

Figure 3 visualizes the geometrical idea behind the SVM. In both pictures, a margin is created by straight parallel lines which separate the two groups. Recall that the distance between the groups needs to be maximized in order to improve classification. If the two groups are perfect linearly separable, meaning that no observations are in the margin or the opposite group as shown on the left picture, then all observations will satisfy

$$\begin{aligned} x_i^\top w + b &\geq 1 \quad \text{for} \quad y_i = 1, \\ x_i^\top w + b &\leq -1 \quad \text{for} \quad y_i = -1, \end{aligned}$$

However, if there are observations in the marginal zone or in the opposite group shown on the right panel of Figure 3, the misclassification needs to be penalized with η. Hence, the inequalities are adjusted for all observations and we get:

$$x_i^\top w + b \geq 1 - \xi_i \quad \text{for} \quad y_i = 1, \tag{8}$$

$$x_i^\top w + b \leq -1 + \xi_i \quad \text{for} \quad y_i = -1, \tag{9}$$

$$\xi_i \geq 0, \tag{10}$$

One of the main advantages of the SVM method is that only the data points on the margin, in the margin and in the opposite group, are used for classification. These observations are then called support vectors giving the technique its name. The logit and DA methods use all observations for classification causing a higher cost of iteration.

From the above inequalities (8) and (9), the primal minimization problem of the SVM

is written as:

$$\min_{w} \frac{1}{2}\|w\|^2 + C\sum_{i=1}^{n}\xi_i, \tag{11}$$

subject to

$$y_i(x_i^\top w + b) \geq 1 - \xi_i \tag{12}$$

$$\xi_i \geq 0. \tag{13}$$

C is the complexity or capacity. Smaller C's lead to larger margins and avoid overfitting, but the misclassification rate is potentially higher. Sometimes each group has a different C. These C's are only necessary if the sizes of both groups are sufficiently different when training the SVM. Furthermore, the second term of (11) serves as the misclassification penalty.

The primal problem above cannot be solved directly. The introduction of Lagrangian multipliers to (11) - (13) leads to following equations:

$$\min_{w,b,\xi_i}\max_{\alpha_i,\mu_i} L_P = \frac{1}{2}\|w\|^2 + C\sum_{i=1}^{n}\xi_i - \sum_{i=1}^{n}\alpha_i\{y_i\left(x_i^\top w + b\right) - 1 + \xi_i\} - \sum_{i=1}^{n}\beta_i\xi_i,$$

where $\alpha_i \geq 0$ and $\beta_i \geq 0$ are the Lagrange multipliers. They are non-zero for support vectors.

Equation (14) needs to be rewritten as a dual problem using the Karush-Kuhn-Tucker conditions (Gale, Kuhn & Tucker 1951). The dual problem is

$$\max_{\alpha_i} = \sum_{i=1}^{n}\alpha_i - \frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{n}\alpha_i\alpha_j y_i y_j x_i^\top x_j, \tag{14}$$

$$\text{s.t.} 0 \leq \alpha_i \leq C, \tag{15}$$

$$\sum_{i=1}^{n}\alpha_i y_i = 0. \tag{16}$$

Note that this optimization problem is convex; therefore, a unique solution can be found.

This dual optimization problem can be solved by hand or with a statistical computer program. Its solutions are the n Lagrangian multipliers α_i. They determine the degree of influence of each training observation. The harder an observation is to classify the higher α_i will be. This explains why the α_i's are zero for the observations lying in the correct area. Once α_i's are calculated the weight w of the d-variables is given:

$$w = \sum_{i=1}^{n}\alpha_i y_i x_i.$$

Notice the Lagrangian multipliers are directly related to the weights. In logistic regression this comparison of α and w is not possible.

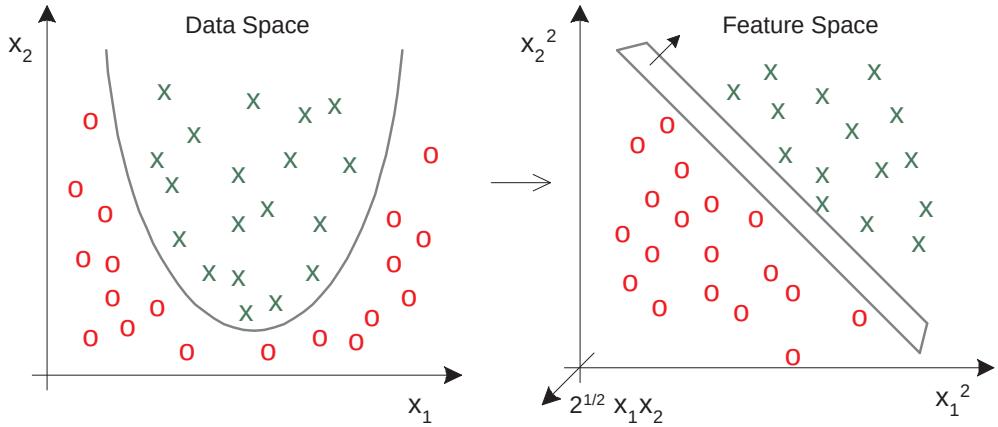

Figure 4: Mapping from a two-dimensional data space into a three-dimensional space of features $\mathbb{R}^2$ to $\mathbb{R}^3$

If the data used is non-linear separable we transform it into a higher dimensional space, called the feature space. The minimization (11) depends only on the scalar product $x^\top x_i$, not on the original x and x_i. Therefore, $x^\top x_i$ can be replaced by a function $k(x, x_i)$ so that there is a mapping from a lower dimensional space into higher dimensional space in which the data is linearly separable. The kernel function $k(x_i, y_j)$ must satisfy the Mercer conditions (Mercer (1909)):

1. $k(x_i, y_j) = k(y_j, x_i) =$ (symmetric)
2. $\forall (x_1, \ldots, x_n), (y_1, \ldots, y_n), k_{ij} = k(x_i, y_j), a^\top k a \geq 0$ (semi-positive definite)

Hence, the data can be mapped into infinitely dimensional spaces as in the case with the Gaussian kernels.

Figure 4 visualizes a simple example of the above theory. Let k be the quadratic kernel function:

$$k(x, x_i) = \left(x^\top x_i\right)^2,$$

which maps from $\mathbb{R}^2$ into $\mathbb{R}^3$. The explicit map is:

$$\psi(x, x_i) = \left(x^2, \sqrt{2} x x_i, x_i^2\right)^2$$

Fortunately, we do not need to determine the transformation ψ explicitly because

$$k(x, x_i) = \psi(x, x_i)^\top \psi(x, x_i).$$

Non-linear data can be classified into a higher dimensional feature space without changing the SVM solution. Higher dimensionality of the data and degree of the polynomial

of k results in a larger number of features. The advantage of the 'kernel trick' becomes obvious: We train the SVM in a higher dimensional feature space, where the data is linear separable, but the feature space does not need to be determined explicitly.

Some examples of kernel functions are:

- $k(x_i, x_j) = (x_i \cdot x_j + c)^d$ (Polynomial)
- $k(x_i, x_j) = \exp\left\{-\left\|x_i - x_j\right\|^2 / \left(2\sigma^2\right)\right\}$ (RBF)
- $K(x_i, x_j) = \tanh(kx_i^\top x_j - \delta)$ – the hyperbolic tangent kernel
- $K(x_i, x_j) = e^{-(x_i - x_j)^\top r^{-2} \Sigma^{-1} (x_i - x_j)/2}$ (stationary Gaussian kernel)

The last kernel has an anisotropic radial basis. Later, we will apply this kernel taking Σ equal to the variance matrix of the training set and r as a constant standing for the radius of the smallest sphere containing data. The higher r the lower the complexity.

Once, the support vector machine is trained and we have found the values for w we are able to classify a new company described by variables of x using the classification rule:

$$g(x) = \text{sign}\left(x^\top w + b\right), \tag{17}$$

where $w = \sum_{i=1}^n \alpha_i y_i x_i$ and $b = \frac{1}{2}\left(x_{+1} + x_{-1}\right) w$. x_{+1} and x_{-1} are any two support vectors belonging to different classes. They must both lay on the margin boundary. To reduce numerical errors when training the SVM it is desirable to use averages over all x_+ and x_- instead of two arbitrarily chosen support vectors. The value of the classification function, or in our study, the score of a company is computed as:

$$f(x) = x^\top w + b. \tag{18}$$

We note that each value of $f(x)$ uniquely corresponds to a default probability (PD).

There are different ways of calculating the PD values for each company. Here, one solution is presented with two steps. First, we calculate for all $i = 1, 2, \ldots, n$ observations of the training set:

$$\tilde{PD}(z) = \frac{\sum_{i=1}^n w(z - z_i) I(y_i = 1)}{\sum_{i=1}^n w(z - z_i)},$$

where $w(z - z_i) = exp\left\{(z - z_i)^2 / 2h^2\right\}$. $z_i = Rank f(x_i)$ is the rank of the ith company. Higher scores $f(x_i)$ lead to higher ranks. The smoothness of $\tilde{PD}(z)$ is influenced by the bandwidth h. Smaller h's give higher smoothness. Using the company rank z_i instead of the score $f(x_i)$ we obtain a $k - NN$ smoother with Gaussian weights $\frac{w(z - z_i)}{\sum_{j=1}^n w(z - z_j}$ which decay gradually as $|z - z_i|$ grows. After the first step, the PD values are not necessarily monotone as the thin black line shows in Figure 5. Therefore, a second step is necessary.

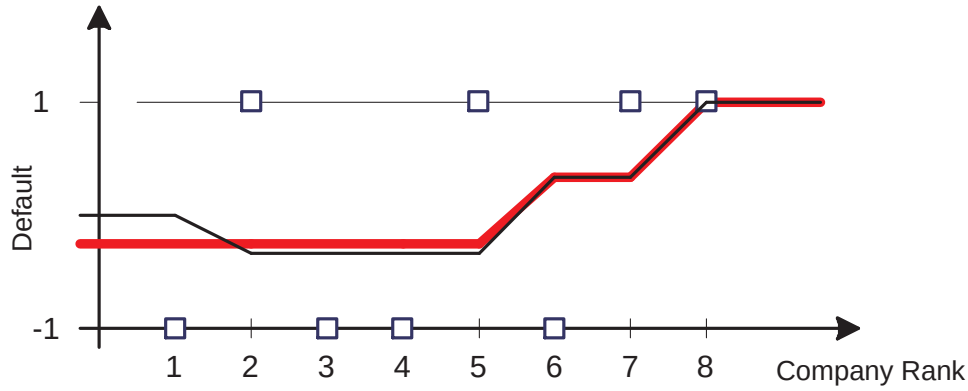

Figure 5: Monotonisation of PD's with the pool adjacent violator algorithm. The thin line denotes PD's estimated with the $k - NN$ method with uniform weights and $k = 3$ before monotonisation and the bold red line after monotonisation. Here $y = 1$ for insolvencies, $y = -1$ for solvent companies

In step 2, we monotonise $\tilde{PD}(z_i)$ by using the Pool Adjacent Violator (PAV) algorithm (Barlow, Bartholomew, Bremmer, and Brunk (1972)). Figure 5 illustrates how the previous black line (before monotonisation) changes to the thick red line (PD after monotonisation). The PD values become monotone. The companies are ordered according to their rank on the horizontal axis. The value $y = 1$ indicates insolvency and $y = -1$ indicates solvency. Between the ranking 1 and 2 monotonicity is violated. After applying the the PAV algorithm, the PD value is corrected.

With the PAV we obtain monotonised probabilities of default, PD(x_i), for the observations of the training set as seen in Figure 6. A PD for any observation x of the testing set is computed by interpolating PDs for two adjacent observations in terms of the score from the training set. If the score for x lies beyond the range of the scores of the training set, then PD(x) is set equal to the score of the first neighbouring observation of the training set.

Finally, we measure the quality of the rating method for which the accuracy ratio (AR) is applied. AR values close to one indicate good performances of the method. If the two groups are the same size then we calculate the AR:

$$AR \approx 2 \int_0^1 y(x)dx - 1.$$

From the AR, we can draw a receiver operating characteristics (ROC) curve as shown in Figure 7. The blue line demonstrates the perfect separation. Due to the intermesh

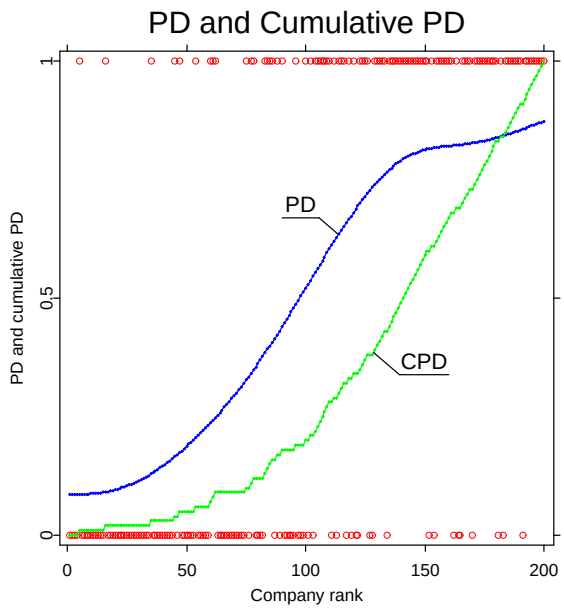

Figure 6: The circles represent the smoothing and monotonisation of default ($y = 1$) and non-default ($y = 0$) companies.

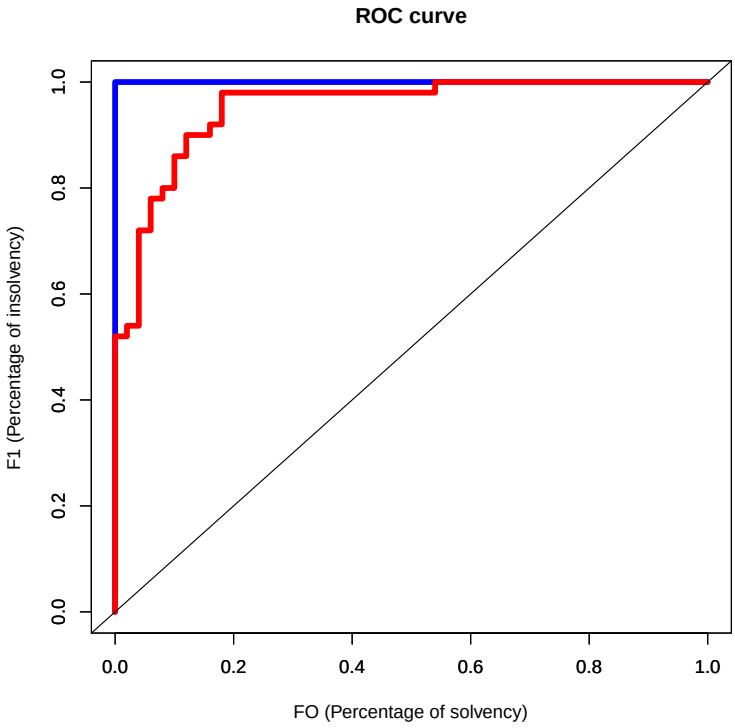

Figure 7: Receiver Operating Characteristics Curve. An SVM is applied with the radial basis $2\Sigma^{1/2}$ and capacity $C = 1$.

of insolvent and solvent companies perfect separation does not exist. Usually the curve looks like the red line. In the case of a naive model, the ROC curve is simply the bisector.

3 Importance of risk classification and Basel II

When we look back at the history of financial markets, we observe many ups and downs. The wish to create a stable financial system will always exist. In 1974, one of the largest post World War II bankruptcies happened, the liquidation of I. D. Herstatt KGaA, a private bank in Cologne. Consequently, a committee was formed with the intent of setting up regulations for banks. One way to prevent bankruptcies is to correctly assert the financial situation of the contractor. In 1988, the Basel Committee on Banking Supervision (BCBS) presented the Basel I Accord. It suggested a minimal capital requirement for banks to prevent future bank crashes.

This treaty did not prevent the world market from a crisis in 2002. Economists realized

that credit rating systems, which developed in the 1950, did not accurately price risks. Others said that the solution to market crashes are efficient regulations. The BCBS established Basel II, which addressed credit and operational risk, and also integrated supervisory review and market discipline. It gives banks the option to use their own rating systems, but they must base the validity of their systems on statistical models. This law came into effect at the end of 2006.

The committee in Basel also decided to require banks to use rating systems with new particular properties such as statistical empirical justification for the calculation of PD values. All the additional requirements will be in effect by 2012. Thus the German Bundesbank, for example, is working hard adjusting their rating software. Unfortunately, up until five years ago, only a small percentage of firms and banks were rated so that developing any classification system brings many challenges. Even today, many companies are not listed on the database. Often, the premium for credits is not based on models but on the sole decision of a loan officer. According to Basel II, a company is declared insolvent if it overdraws its debts 90 days after maturity.

Lending arbitrarily is not safe, and it is better to calculate the default probability (PD) of the borrower. The PD value can be directly calculated from the score found with the SVM method as explained in the previous section. Then, one can decide to which rating class the borrower belongs as shown in Table 1 or in Table 2 Those two tables are just general examples but do not correspond to each industry because different tables for different countries and industries exist. Banks and insurances can use tables such as 1 and 2 to establish the risk premium for each class. Whereas Table 3 displays the capital requirements according to Basel I and Basel II.

4 Description of data

The dataset used in this work comes from the credit reform database provided by the Research Data Center (RDC) of the Humboldt Universität zu Berlin. It contains financial information from 20000 solvent and 1000 insolvent German companies. The time period ranges from 1996 to 2002 and in the case of the insolvent companies the information was gathered 2 years before the insolvency took place. The last annual report of a company before it went bankrupt receives the indicator $y = 1$ and for the rest (solvent) $y = -1$.

We are given 28 variables, i.e. cash, inventories, equity, EBIT, number of employees and branch code. From the original data, we create common financial indicators which are denoted as $x1 \dots x25$. These ratios can be grouped into four categories such as profitability, leverage, liquidity and activity. The indicators are presented in Table 4. For the $x9$ formula, INGA and LB mean intangible assets and lands & buildings, respectively.

Rating Class (S&P)	One year PD (%)
AAA	0.01
AA+	0.01
AA	0.02
AA-	0.03
A+	0.05
A	0.08
A-	0.13
BBB+	0.22
BBB	0.36
BBB-	0.58
BB+	0.94
BB	1.55
BB-	2.50
B+	4.08
B	6.75
B-	10.88
CCC+	17.75
CCC-	29.35
D	100.00

Table 1: Rating classes and PDs. Source: Henking (2006)

Rating Class (S&P)	Five year PD (%)
AAA	0.08
AA	0.32
A	0.91
A	0.08
BBB	3.45
BB	12.28
B	32.57
CCC-	69.75
D	100.00

Table 2: Rating classes and PDs. Henking (2006)

Rating Class (S&P)	One-year PD (%)	Capital Requirements (%) (Basel I)	Capital Requirements (%) (Basel II)
AAA	0.01	8.00	0.63
AA	0.02 – 0.04	8.00	0.93 – 1.40
A+	0.05	8.00	1.60
A	0.08	8.00	2.12
A-	0.11	8.00	2.55
BBB	0.15 – 0.40	8.00	3.05 – 5.17
BB	0.65 – 1.95	8.00	6.50 – 9.97
B+	3.20	8.00	11.90
B	7.00	8.00	16.70
B-	13.00	8.00	22.89
CCC	> 13	8.00	> 22.89
CC		8.00	
C		8.00	
D		8.00	

Table 3: Rating grades and capital requirements. Source: (Damodaran, 2002) and (Füser, 2002). The figures in the last column were estimated by the authors for a loan to an SME with a turnover of 5 million Euros with a maturity of 2.5 years using the data from column 2 and the recommendations of the Basel Committee on Banking Supervision (BCBS, 2003).

Ratio No.	Definition	Ratio	Category
$x1$	NI/TA	Return on assets	Profit.
$x2$	NI/Sales	Net profit margin	Profit.
$x3$	OI/TA	Operating Inc./Total ass.	Profit.
$x4$	OI/Sales	Operating profit margin	Profit.
$x5$	EBIT/TA	EBIT/Total assets	Profit.
$x6$	(EBIT+AD)/TA	EBITDA	Profit.
$x7$	EBIT/Sales	EBIT/Sales	Profit.
$x8$	Equity/TA	Own funds ratio (simple)	Leverage
$x9$	(Equity-ITGA)/ (TA-ITGA-Cash-LB)	Own funds ratio (adj.)	Leverage
$x10$	CL/TA	Current liab./Total ass.	Leverage
$x11$	(CL-Cash)/TA	Net indebtedness	Leverage
$x12$	TL/TA	Total liab./Total ass.	Leverage
$x13$	Debt/TA	Debt ratio	Leverage
$x14$	EBIT/Interest exp.	Interest coverage ratio	Leverage
$x15$	Cash/TA	Cash/Total assets	Liquidity
$x16$	Cash/CL	Cash ratio	Liquidity
$x17$	QA/CL	Quick ratio	Liquidity
$x18$	CA/CL	Current ratio	Liquidity
$x19$	WC/TA	Working Capital	Liquidity
$x20$	CL/TL	Current liab./Total liab.	Liquidity
$x21$	TA/Sales	Asset turnover	Activity
$x22$	INV/Sales	Inventory turnover	Activity
$x23$	AR/Sales	Account receiv. turnover	Activity
$x24$	AP/Sales	Account payable turnover	Activity
$x25$	Log(TA)	Log(Total assets)	Activity

Table 4: Defintions of financial ratios.

Ratio	$q_{0.05}$	Med.	$q_{0.95}$	IQR	$q_{0.05}$	Med.	$q_{0.05}$	IQR
$x1$	-0.19	0.00	0.09	0.04	-0.09	0.02	0.19	0.06
$x2$	-0.015	0.00	0.06	0.03	-0.07	0.01	0.10	0.03
$x3$	-0.22	0.00	0.10	0.06	-0.011	0.03	0.27	0.09
$x4$	-0.16	0.00	0.07	0.04	-0.08	0.02	0.13	0.04
$x5$	-0.19	0.02	0.13	0.07	-0.09	0.05	0.27	0.09
$x6$	-0.13	0.07	0.21	0.08	-0.04	0.11	0.35	0.12
$x7$	-0.14	0.01	0.10	0.04	-0.07	0.02	0.14	0.05
$x8$	0.00	0.05	0.40	0.13	0.00	0.14	0.60	0.23
$x9$	-0.01	0.05	0.56	0.17	0.00	0.16	0.95	0.32
$x10$	0.18	0.52	0.91	0.36	0.09	0.42	0.88	0.39
$x11$	-0.12	0.49	0.89	0.36	-0.05	0.36	0.83	0.41
$x12$	0.29	0.76	0.98	0.35	0.16	0.65	0.96	0.40
$x13$	0.00	0.21	0.61	0.29	0.00	0.15	0.59	0.31
$x14$	-7.90	1.05	7.20	2.47	-6.78	2.16	73.95	5.69
$x15$	0.00	0.02	0.16	0.05	0.00	0.03	0.32	0.10
$x16$	0.00	0.03	0.43	0.11	0.00	0.08	1.40	0.29
$x17$	0.18	0.68	1.90	0.54	0.25	0.94	4.55	1.00
$x18$	0.56	1.26	3.73	0.84	0.64	1.58	7.15	1.56
$x19$	-0.32	0.15	0.63	0.36	-0.22	0.25	0.73	0.41
$x20$	0.34	0.84	1.00	0.37	0.22	0.85	1.00	0.4
$x21$	0.43	1.63	4.15	1.41	0.50	2.08	6.19	1.76
$x22$	0.02	0.16	0.89	0.26	0.01	0.11	0.56	0.16
$x23$	0.02	0.12	0.33	0.11	0.00	0.09	0.25	0.09
$x24$	0.03	0.14	0.36	0.10	0.01	0.07	0.24	0.08
$x25$	13.01	14.87	17.16	1.69	12.82	17.95	1.657	2.37

Table 5: Descriptive statistics for financial ratios.

5 Calculations

In order to reduce the effect of the outliers on the results, all observations that exceeded the upper limit of $Q75+1.5*IQ$ (Inter-quartile range) or the lower limit of $Q25-1.5*IQ$ were replaced with these values. Table 5 gives an overview of the summary statistics. In the next table (6), the insolvent and solvent companies for each year are displayed. Insolvent company data for the year of 1996 are missing, we will therefore exclude them from further calculations. We are left with 1000 insolvent and 18610 solvent companies.

Not all variables are good predictors for the classification method. The most common techniques to find the right variables are Mallows' CP, backward stepwise and forward stepwise selection. The latter is preferred when dealing with many variables because the cost of computation is reduced. This method starts with a univariate model and

Year	Solvent	Insolv.	Solv. Ratio	Insolv. Ratio
1996	1390	0	100.00%	0.00%
1997	1468	146	90.95%	9.05%
1998	1615	1274	92.71%	7.29%
1999	2124	179	92.23%	7.77%
2000	3086	175	94.63%	5.37%
2001	4380	187	95.91%	4.09%
2002	5937	186	96.96%	3.04%

Table 6: The distribution of the data over the years for solvent and insolvent companies.

continues adding variables until all variables are included. At each step the variable is kept whose addition to the model resulted in the highest median accuracy ratio (AR). As a result, ratios $x24, x3, x15, x12, x25, x22, x5$, and $x2$ generate the model with the highest AR (60.51%).

6 Computational results

After choosing the best predictors, we start calculating score values for each company. As mentioned earlier, the results depend on the predefined C, the capacity, and r. To demonstrate how performance changes, we will use the accounts payable turnover ($x24$), which is the best univariate model, and calculate Type I and Type II errors. Type I errors are those when companies were predicted to stay solvent but turned insolvent. Type II errors are the mistakes we make by assuming companies will default but they do not. Keep in mind that different kernels will also influence performance. We use one of the most common ones, the radial Gaussian kernel.

In Figures 8–11 the triangles represent solvent and circles represent insolvent companies from a chosen training set. The solid shapes represent the support vectors. We randomly chose 50 solvent and 50 insolvent companies. The colored background corresponds to different score values f. The bluer the area, the higher the score and probability of default. Most successful companies are in the red area and have positive profitability and reasonable activity.

Figure 8 presents the classification results for an SVM using $r = 100$ and the fixed capacity $C = 1$. With the given priors, the SVM has trouble classifying between solvent and insolvent companies. The radial base r, which determines the minimum radius of a group, is too large. Default companies do not seem to exist. Notice that the SVM is doing a poor job of distinguishing between the groups even though most observations are used as support vectors.

C	r	Type I error	Type II error
0.001	0.600	40.57	23.43
0.100	0.600	38.42	24.45
10.000	0.600	34.43	27.86
100.000	0.600	25.22	34.44
1000.000	0.600	25.76	34.26
10.0	0.002	37.20	32.79
10.0	0.060	31.86	29.25
10.0	6.000	36.97	27.86
10.0	60.000	37.27	25.87
10.0	2000.000	41.09	24.85

Table 7: Misclassification errors (30 randomly selected samples; one predictor $x24$).

In Figure 9, the minimal radius is reduced to 2 while C remains the same. Clearly, the SVM starts recognizing the difference between solvent and insolvent companies resulting in sharper clusters for successful and failing companies. Decreasing r even further, i.e. to 0.5 as in Figure 10, emphasizes the groups in even more detail. If the radial base is too small, then the complexity will be too high for a given data set.

If we increase the capacity C, we decrease the distance between the groups. Figure 11 demonstrates this effect on the classification results. The beige area outside the clusters is associated with score values of around zero. With higher C's, the SVM localizes only one cluster of successful companies. It is crucial to use statistical methods, i.e. the leave-one-out method, to find the optimal priors C and r for classifying companies.

After a SVM classification function f is trained, we can calculate the scores for new companies, determine their PD values and decide if they belong to the solvent or insolvent group. Further, a SVM learns the cluster of both groups given that the constants C and r are chosen appropriately. If the capacity is too high the knowledge of cluster centre vanishes. If r is too high, the groups might intermesh. The choice of the kernel k also affects the solution. Often, the most appropriate kernel is one of the Gaussian kernel.

Earlier we introduced one way of converting scores into PDs including a smoothing technique. However, we can also calibrate the PDs by hand. Consider an example with $r = 2$ and $C = 1$. We choose three rating classes: safe, neutral and risky and give each class a corresponding score value of $f < -0.0115$, $-0.0115 < f < 0.0115$ and $f > 0.0115$, respectively. Next we count how many companies belong to each group then calculate the ratio of failing companies giving us the estimated probability of default for each rating class.

With a sufficient number of observations in the training set, the rating classes could be divided up into finer ones. We have seen an example in Table 1. The rating company S

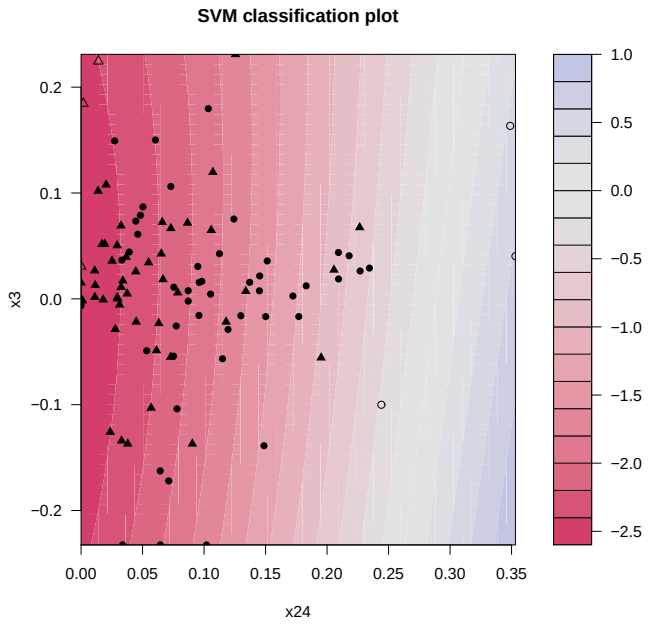

Figure 8: Ratings of companies in two dimensions. The case of a low complexity of classifier functions, the radial basis is 100, the capacity is fixed at $C = 1$. STF2svm01

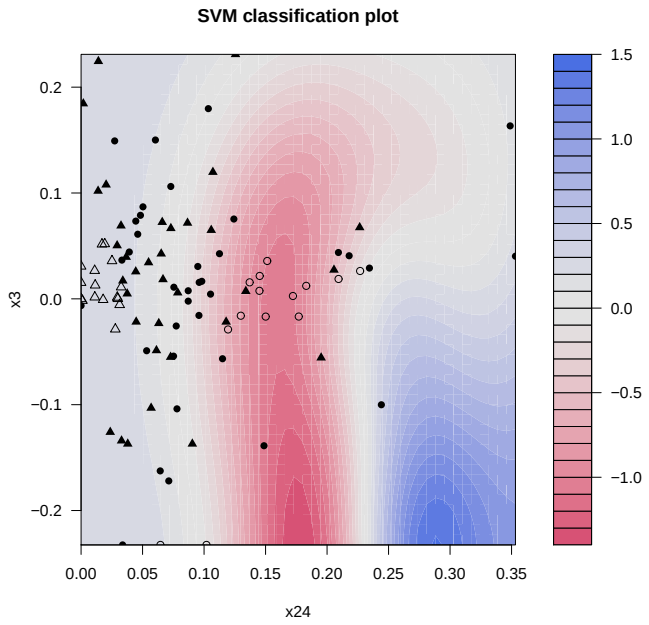

Figure 9: Ratings of companies in two dimensions. The case of an average complexity of classifier functions, the radial basis is 2, the capacity is fixed at $C = 1$. STF2svm01

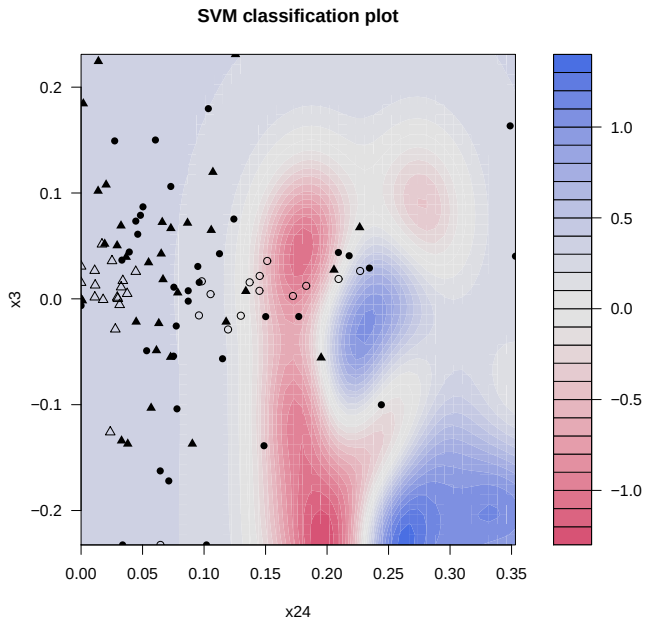

Figure 10: Ratings of companies in two dimensions. The case of an excessively high complexity of classifier functions, the radial basis is 0.5, the capacity is fixed at $C = 1$. STF2svm01

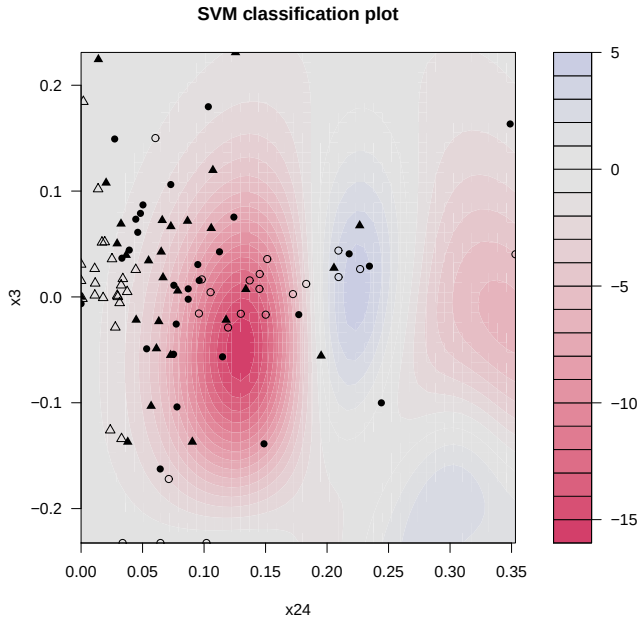

Figure 11: Ratings of companies in two dimensions, the case of a high capacity ($C = 200$). The radial basis is fixed at 2. STF2svm01

& P uses over a dozen rating classes such as AAA, AA ,. . . Different rating agencies use a variety of numbers and names for their classifications.

{Conclusions

SVMs are capable of predicting whether a company will be solvent or insolvent based only on its financial and economic information. Of course, the SVM needs to be provided with a training data set. Once we have learned the SVM, it divulges the financial situation of a company, which may not be obvious at first glance. SVMs are easy to implement with their low number of calculational steps and priors. They often give the best classification results compared to logit or other methods. Thus, SVMs have become more and more popular over the last decade.

We have learned that the scores found with SVM models can be used to calculate the individual PDs for each company. Consequently, credits and other financial instruments can be adjusted accordingly. Banks and companies will profit from those results because it helps them to decide with what kind of risk they can carry. This leads hopefully to a more stable financial market by increasing our ability to predict defaults and accurately evaluate risk.

References

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

Beaver, W. (1966). Financial ratios as predictors of failures. empirical research in accounting: Selected studies, *Journal of Accounting Research* pp. 71–111. supplement to vol. 5.

Dimitras, A., Slowinski, R., Susmaga, R. & Zopounidis, C. (1999). Business failure prediction using rough sets, *European Journal of Operational Research* (114): 263–280.

Eunite (2001). Electricity load forecast competition of the EUropean Network on Intelligent TEchnologies for Smart Adaptive Systems, *http://neuron.tuke.sk/competition/* .

Fitzpatrick, P. (1932). A comparison of the ratios of successful industrial enterprises with those of failed companies.

Frydman, H., Altman, E. & Kao, D.-L. (1985). Introducing recursive partitioning for financial classification: The case of financial distress, *The Journal of Finance* **40**(1): 269–291.

Gale, D., Kuhn, H. W. & Tucker, A. W. (1951). *Linear Programming and the Theory of Games, in Activity Analysis of Production and Allocation, T. C. Koopmans (ed.)*, John Wiley & Sons, New York, NY.

Henking, A. (2006). *Kreditrisikomessung*, first edn, Springer.

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

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

Ramser, J. & Foster, L. (1931). A demonstration of ratio analysis. bulletin no. 40.

S. Chen, W. H. a. R. M. (accepted in 2009). Modelling default risk with support vector machines, *Journal of Quantitative Finance* .

Schölkopf, B. (1997). *Support vector learning*, Oldenbourg.

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

Vapnik, V. (1995). *The Nature of Statistical Learning Theory*, Springer, New York.

Wiginton, J. (1980). A note on the comparison of logit and discriminant models of consumer credit behaviour, *Journal of Financial and Quantitative Analysis* **15**(3): 757–770.

Wilcox, A. (1971). A simple theory of financial ratios as predictors of failure, *Journal of Accounting Research* pp. 389–395.

Winakor, A. & Smith, R. (1935). Changes in the financial structure of unsuccessful industrial corporations. bulletin no. 51.

Zavgren, C. (1983). The prediction of corporate failure: The state of the art, *Journal of Accounting Literature* (2): 1–38.

Zmijewski, M. (1984). Methodological issues related to the estimation of financial distress prediction models, *Journal of Accounting Research* **20**(0): 59–82.

SFB 649 Discussion Paper Series 2010

For a complete list of Discussion Papers published by the SFB 649, please visit http://sfb649.wiwi.hu-berlin.de.

001 "Volatility Investing with Variance Swaps" by Wolfgang Karl Härdle and Elena Silyakova, January 2010.
002 "Partial Linear Quantile Regression and Bootstrap Confidence Bands" by Wolfgang Karl Härdle, Ya'acov Ritov and Song Song, January 2010.
003 "Uniform confidence bands for pricing kernels" by Wolfgang Karl Härdle, Yarema Okhrin and Weining Wang, January 2010.
004 "Bayesian Inference in a Stochastic Volatility Nelson-Siegel Model" by Nikolaus Hautsch and Fuyu Yang, January 2010.
005 "The Impact of Macroeconomic News on Quote Adjustments, Noise, and Informational Volatility" by Nikolaus Hautsch, Dieter Hess and David Veredas, January 2010.
006 "Bayesian Estimation and Model Selection in the Generalised Stochastic Unit Root Model" by Fuyu Yang and Roberto Leon-Gonzalez, January 2010.
007 "Two-sided Certification: The market for Rating Agencies" by Erik R. Fasten and Dirk Hofmann, January 2010.
008 "Characterising Equilibrium Selection in Global Games with Strategic Complementarities" by Christian Basteck, Tijmen R. Daniels and Frank Heinemann, January 2010.
009 "Predicting extreme VaR: Nonparametric quantile regression with refinements from extreme value theory" by Julia Schaumburg, February 2010.
010 "On Securitization, Market Completion and Equilibrium Risk Transfer" by Ulrich Horst, Traian A. Pirvu and Gonçalo Dos Reis, February 2010.
011 "Illiquidity and Derivative Valuation" by Ulrich Horst and Felix Naujokat, February 2010.
012 "Dynamic Systems of Social Interactions" by Ulrich Horst, February 2010.
013 "The dynamics of hourly electricity prices" by Wolfgang Karl Härdle and Stefan Trück, February 2010.
014 "Crisis? What Crisis? Currency vs. Banking in the Financial Crisis of 1931" by Albrecht Ritschl and Samad Sarferaz, February 2010.
015 "Estimation of the characteristics of a Lévy process observed at arbitrary frequency" by Johanna Kappusl and Markus Reiß, February 2010.
016 "Honey, I'll Be Working Late Tonight. The Effect of Individual Work Routines on Leisure Time Synchronization of Couples" by Juliane Scheffel, February 2010.
017 "The Impact of ICT Investments on the Relative Demand for High-Medium-, and Low-Skilled Workers: Industry versus Country Analysis" by Dorothee Schneider, February 2010.
018 "Time varying Hierarchical Archimedean Copulae" by Wolfgang Karl Härdle, Ostap Okhrin and Yarema Okhrin, February 2010.
019 "Monetary Transmission Right from the Start: The (Dis)Connection Between the Money Market and the ECB's Main Refinancing Rates" by Puriya Abbassi and Dieter Nautz, March 2010.
020 "Aggregate Hazard Function in Price-Setting: A Bayesian Analysis Using Macro Data" by Fang Yao, March 2010.
021 "Nonparametric Estimation of Risk-Neutral Densities" by Maria Grith, Wolfgang Karl Härdle and Melanie Schienle, March 2010.

SFB 649 Discussion Paper Series 2010

For a complete list of Discussion Papers published by the SFB 649, please visit http://sfb649.wiwi.hu-berlin.de.

022 "Fitting high-dimensional Copulae to Data" by Ostap Okhrin, April 2010.
023 "The (In)stability of Money Demand in the Euro Area: Lessons from a Cross-Country Analysis" by Dieter Nautz and Ulrike Rondorf, April 2010.
024 "The optimal industry structure in a vertically related market" by Raffaele Fiocco, April 2010.
025 "Herding of Institutional Traders" by Stephanie Kremer, April 2010.
026 "Non-Gaussian Component Analysis: New Ideas, New Proofs, New Applications" by Vladimir Panov, May 2010.
027 "Liquidity and Capital Requirements and the Probability of Bank Failure" by Philipp Johann König, May 2010.
028 "Social Relationships and Trust" by Christine Binzel and Dietmar Fehr, May 2010.
029 "Adaptive Interest Rate Modelling" by Mengmeng Guo and Wolfgang Karl Härdle, May 2010.
030 "Can the New Keynesian Phillips Curve Explain Inflation Gap Persistence?" by Fang Yao, June 2010.
031 "Modeling Asset Prices" by James E. Gentle and Wolfgang Karl Härdle, June 2010.
032 "Learning Machines Supporting Bankruptcy Prediction" by Wolfgang Karl Härdle, Rouslan Moro and Linda Hoffmann, June 2010.