Variable Selection and Optimization in Default Prediction

Dedy Dwi Prastyo
Wolfgang Karl Härdle

Ladislaus von Bortkiewicz Chair of Statistics
C.A.S.E. – Center for Applied Statistics and Economics
Humboldt-Universität zu Berlin
http://lvb.wiwi.hu-berlin.de
http://www.case.hu-berlin.de
Figure 1: Who is more precise?
Credit rating

- **Score** $(S)$
  Quantitative indicator for customers w.r.t. their individual default risk

- **Probability of Default** (PD)
  One-to-one mapping of the score, $S \rightarrow PD(S)$

- **Rating**
  Classification of customers (private, corporate, sovereign) into groups of equivalent default risk
Introduction

Default prediction

- **Time-series data (market data)**
  - Merton approach (stock price as estimate for the market value), \( S = \text{distance to default} \)

- **Cross-sectional data (i.e. balance sheet)**
  - Discriminant analysis
  - Categorical regression (logit, probit)
  - Support Vector Machines (SVM)
Research questions

- What are the variables (i.e. accounting) significantly contribute to default?
- How to optimize the default prediction (classification)?
Outline

1. Introduction ✓
2. Variable Selection – Regularized GLM, elastic net
3. Evolutionary optimization
4. Genetic Algorithm SVM
Some problems

- Number of predictors is greater than number of observation, $p \gg n$
- There are correlated variables
- Sparsity (elements of predictor matrix $X \approx 0$)
- VLDS (very large data set)
Model with convex penalty

Apply fast algorithm to estimate

- Model
  - Linear regression
  - Two-class logistic regression

- Penalties
  - Lasso ($\ell_1$)
  - Ridge regression ($\ell_2$)
  - Elastic net (mixture of $\ell_1$ and $\ell_2$)
Linear regression

Suppose $Y \in \mathbb{R}$ and $X \in \mathbb{R}^p$

$$E(Y|X = x) = \beta_0 + x^\top \beta,$$

If $\theta = (\beta_0, \beta)$, a penalty $P_\alpha(\beta)$ and multiplier $\lambda$, then

$$\hat{\theta} = \argmin \left\{ \frac{1}{2n} \sum_{i=1}^{n} \left( y_i - \beta_0 - x_i^\top \beta \right)^2 + \lambda P_\alpha(\beta) \right\}$$  \hspace{1cm} (1)

Default Prediction
Elastic net

The penalty is a compromise between ridge and lasso

\[ P_\alpha(\beta) = \frac{1}{2}(1 - \alpha) \|\beta\|_2^2 + \alpha \|\beta\|_1 \]

\[ = \sum_{j=1}^{p} \left\{ \frac{1}{2}(1 - \alpha)\beta_j^2 + \alpha|\beta_j| \right\} \quad \text{(2)} \]

Be a ridge regression, if \((\alpha = 0)\), and lasso, if \((\alpha = 1)\).

Useful when \(p \gg n\) or there are many correlated variables.
Figure 2: Profile of estimated coefficients, $\lambda = 1, \ldots, 10$, elastic net ($\alpha = 0.2$), on Leukemia data, $p = 72$ and $n = 3571$ (Friedman et al., 2010)
Binary logit

Suppose \( p(x_i) = P(Y = 1|x_i) \), with \( Y \in \{0, 1\} \),

\[
P(Y = 1|x) = \left\{ 1 + e^{-(\beta_0 + x^T \beta)} \right\}^{-1},
\]

\[
P(Y = 0|x) = \left\{ 1 + e^{(\beta_0 + x^T \beta)} \right\}^{-1},
\]

\[
\log \left\{ \frac{P(Y = 1|x)}{P(Y = 0|x)} \right\} = \beta_0 + x^T \beta
\]
Penalized log likelihood

\[
\max_{\beta_0, \beta} \left\{ \frac{1}{n} \sum_{i=1}^{n} \ell(\beta_0, \beta) - \lambda P_\alpha(\beta) \right\}
\]

(4)

where

\[
\ell(\beta_0, \beta) = \frac{1}{n} \sum_{i=1}^{n} y_i (\beta_0 + x_i^T \beta) - \log(1 + e^{\beta_0 + x_i^T \beta}),
\]

(5)

is a concave function of the parameter. Maximizing \( \ell(\beta_0, \beta) \) — iteratively reweighted least squares (IRLS).
Newton algorithm

If current parameter are \((\tilde{\beta}_0, \tilde{\beta})\), the quadratic approximation to \(\ell(\beta_0, \beta)\) is,

\[
\ell_Q(\beta_0, \beta) = -\frac{1}{2n} \sum_{i=1}^{n} w_i (z_i - \beta_0 - x_i^\top \beta)^2 + C(\tilde{\beta}_0, \tilde{\beta})^2
\]  

(6)

where working response and weight are,

\[
\begin{align*}
z_i &= \tilde{\beta}_0 + x_i^\top \tilde{\beta} + \frac{y_i - \tilde{p}(x_i)}{\tilde{p}(x_i)(1 - \tilde{p}(x_i))} \\
w_i &= \tilde{p}(x_i)(1 - \tilde{p}(x_i))
\end{align*}
\]

Newton update is obtained by minimizing \(\ell_Q(\beta_0, \beta)\).
Friedman approach

Coordinate descent is used to solve the penalized weighted least-square (PWLS)

$$\min_{\beta_0, \beta} \{-\ell_Q(\beta_0, \beta) + \lambda P_\alpha(\beta)\}$$

Sequence of nested loops:
- Outer loop: Decrement $\lambda$
- Middle loop: Update $\ell_Q$ using current parameter $(\tilde{\beta}_0, \tilde{\beta})$
- Inner loop: Run coordinate descent algorithm on PWLS
Global optimum

- Coordinate descent search local minimum
- How to choose $\alpha$ (and $\lambda$)?
- More complicated
Evolutionary optimization

- Genetic Algorithm (GA)
- GA finds global optimum solution – parameters
What is a Genetic Algorithm?

Genetics algorithm is search and optimization technique based on Darwin’s principle on natural selection (Holland, 1975)
Classifier

Figure 3: Linear classifier functions (1 and 2) and a non-linear one (3)
SVM

Classification
Data $D_n = \{(x_1, y_1), \ldots, (x_n, y_n)\} : \Omega \rightarrow (\mathcal{X} \times \mathcal{Y})^n$
$\mathcal{X} \subseteq \mathbb{R}^d$ and $\mathcal{Y} \in \{-1, 1\}$

Goal – to predict $\mathcal{Y}$ for new observation, $x \in \mathcal{X}$, based on information in $D_n$
Linearity (Non-) Separable Case

Figure 4: Hyperplane and its margin in linearly (non-) separable case
## Loss function

<table>
<thead>
<tr>
<th>$L(y, f(x))$</th>
<th>Loss type</th>
</tr>
</thead>
<tbody>
<tr>
<td>${1 - yf(x)}^2$</td>
<td>quadratic loss (ridge regression)</td>
</tr>
<tr>
<td>${1 - yf(x)}_+ = \max{0, 1 - yf(x)}$</td>
<td>hinge loss (SVM)</td>
</tr>
<tr>
<td>$\log{1 + \exp(-yf(x))}$</td>
<td>log-loss (logistic regression)</td>
</tr>
<tr>
<td>$\text{sign}{-yf(x)}$</td>
<td>${0, 1}$ loss</td>
</tr>
<tr>
<td>$1/{1 + \exp(yf(x))}$</td>
<td>sigmoidal loss</td>
</tr>
</tbody>
</table>

Table 1: Types of Loss function, with $f(x) = x^T w + b$ is a score
Figure 5: Plot of loss function for $y = 1$, $f(x) = 0.2x_1$, $x_1 \in [-5, 10]$. Similar plot for $y = -1$ but in opposite direction of $yf(x)$ axis
**SVM dual problem**

\[
\max_{\alpha} L_D (\alpha) = \max_{\alpha} \left\{ \sum_{i=1}^{n} \alpha_i - \frac{1}{2} \sum_{i=1}^{n} \sum_{j=1}^{n} \alpha_i \alpha_j y_i y_j x_i^\top x_j \right\},
\]

s.t. \[0 \leq \alpha_i \leq C\]

\[\sum_{i=1}^{n} \alpha_i y_i = 0\]
Figure 6: Mapping two dimensional data space into a three dimensional feature space, $\mathbb{R}^2 \mapsto \mathbb{R}^3$. The transformation $\Psi(x_1, x_2) = (x_1^2, \sqrt{2}x_1x_2, x_2^2)^T$ corresponds to $K(x_i, x_j) = (x_i^\top x_j)^2$
Non-linear SVM

\[
\max_{\alpha} L_D (\alpha) = \max_{\alpha} \left\{ \sum_{i=1}^{n} \alpha_i - \frac{1}{2} \sum_{i=1}^{n} \sum_{j=1}^{n} \alpha_i \alpha_j y_i y_j K(x_i, x_j) \right\}
\]

s.t. \quad 0 \leq \alpha_i \leq C, \quad \sum_{i=1}^{n} \alpha_i y_i = 0

- Gaussian RBF kernel \(- K(x_i, x_j) = \exp \left(-\frac{1}{\sigma} \|x_i - x_j\|^2\right)\)

- Polynomial kernel \(- K(x_i, x_j) = (x_i^\top x_j + 1)^p\)
Structural Risk Minimization (SRM)

Search for the model structure $S_h$,

$$S_{h_1} \subseteq S_{h_2} \subseteq \ldots \subseteq S_{h^*} \subseteq \ldots \subseteq S_{h_k} = \mathcal{F}$$

such that $f \in S_{h^*}$ minimises the expected risk bound, with $f \subseteq \mathcal{F}$ is class of linear function and $h$ is VC dimension i.e.

$$\text{SVM}(h_1) \subseteq \ldots \subseteq \text{SVM}(h^*) \subseteq \ldots \subseteq \text{SVM}(h_k) = \mathcal{F}$$

with $h$ correspond to the value of SVM (kernel) parameter


Figure 7: Iteration (generation) in GA-SVM
Validation of scores

Discriminatory power (of the score)
- Cumulative Accuracy Profile (CAP) curve
- Receiver Operating Characteristic (ROC) curve
- Accuracy, Specificity, Sensitivity
Figure 8: CAP curve (left) and ROC curve (right)
Discriminatory power

- **Cumulative Accuracy Profile (CAP) curve**
  - CAP/Power/Lorenz curve → Accuracy Ratio (AR)
  - Total sample vs. default sample

- **Receiver Operating Characteristic (ROC) curve**
  - ROC curve → Area Under Curve (AUC)
  - Non-default sample vs. default sample

- Relationship: $\text{AR} = 2 \times \text{AUC} - 1$
# Discriminatory power (cont’d)

<table>
<thead>
<tr>
<th>Sample</th>
<th>Default (1)</th>
<th>Non-default (-1)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Predicted</td>
<td>True Positive (TP)</td>
<td>False Positive (FP)</td>
</tr>
<tr>
<td></td>
<td>False Negative (FN)</td>
<td>True Negative (TN)</td>
</tr>
<tr>
<td>Total</td>
<td>P</td>
<td>N</td>
</tr>
</tbody>
</table>

- **Accuracy**, $P(\hat{Y} = Y) = \frac{TP + TN}{P + N}$

- **Specificity**, $P(\hat{Y} = -1|Y = -1) = \frac{TN}{N}$

- **Sensitivity**, $P(\hat{Y} = 1|Y = 1) = \frac{TP}{P}$
## Credit reform data

<table>
<thead>
<tr>
<th>type</th>
<th>solvent (%)</th>
<th>insolvent (%)</th>
<th>total (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Manufacturing</td>
<td>27.37 (26.06)</td>
<td>25.70 (1.22)</td>
<td>27.29</td>
</tr>
<tr>
<td>Construction</td>
<td>13.88 (13.22)</td>
<td>39.70 (1.89)</td>
<td>15.11</td>
</tr>
<tr>
<td>Wholesale and retail</td>
<td>24.78 (23.60)</td>
<td>20.10 (0.96)</td>
<td>24.56</td>
</tr>
<tr>
<td>Real estate</td>
<td>17.28 (16.46)</td>
<td>9.40 (0.45)</td>
<td>16.90</td>
</tr>
<tr>
<td>total</td>
<td>83.31 (79.34)</td>
<td>94.90 (4.52)</td>
<td>83.86</td>
</tr>
<tr>
<td>others</td>
<td>16.69 (15.90)</td>
<td>5.10 (0.24)</td>
<td>16.14</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>#</th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>20,000</td>
<td>1,000</td>
<td>21,000</td>
</tr>
</tbody>
</table>

Table 2: Credit reform data
## Pre-processing

<table>
<thead>
<tr>
<th>year</th>
<th>solvent # (%)</th>
<th>insolvent # (%)</th>
<th>total # (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1997</td>
<td>872 (9.08)</td>
<td>86 (0.90)</td>
<td>958 (9.98)</td>
</tr>
<tr>
<td>1998</td>
<td>928 (9.66)</td>
<td>92 (0.96)</td>
<td>1020 (10.62)</td>
</tr>
<tr>
<td>1999</td>
<td>1005 (10.47)</td>
<td>112 (1.17)</td>
<td>1117 (11.63)</td>
</tr>
<tr>
<td>2000</td>
<td>1379 (14.36)</td>
<td>102 (1.06)</td>
<td>1481 (15.42)</td>
</tr>
<tr>
<td>2001</td>
<td>1989 (20.71)</td>
<td>111 (1.16)</td>
<td>2100 (21.87)</td>
</tr>
<tr>
<td>2002</td>
<td>2791 (29.07)</td>
<td>135 (1.41)</td>
<td>2926 (30.47)</td>
</tr>
<tr>
<td>total</td>
<td>8964 (93.36)</td>
<td>638 (6.64)</td>
<td>9602 (100)</td>
</tr>
</tbody>
</table>

Table 3: Pre-processed credit reform data
**Full model,** $X_1, \ldots, X_{28}$

- Predictors – 28 financial ratio variables
- Population (# solutions) – 20
- Evolutionary iteration (generation) – 100
- Elitism – 0.2 of population
- Crossover rate – 0.5, mutation rate – 0.1
- Optimal SVM parameters – $\sigma = 1/178.75$ and $C = 63.44$
### Scenario Finding

<table>
<thead>
<tr>
<th>Scenario</th>
<th>training set</th>
<th>testing set</th>
</tr>
</thead>
<tbody>
<tr>
<td>Scenario-1</td>
<td>1997</td>
<td>1998</td>
</tr>
<tr>
<td>Scenario-2</td>
<td>1997-1998</td>
<td>1999</td>
</tr>
<tr>
<td>Scenario-3</td>
<td>1997-1999</td>
<td>2000</td>
</tr>
<tr>
<td>Scenario-4</td>
<td>1997-2000</td>
<td>2001</td>
</tr>
<tr>
<td>Scenario-5</td>
<td>1997-2001</td>
<td>2002</td>
</tr>
</tbody>
</table>

Table 4: Training and testing data set
Quality of classification

<table>
<thead>
<tr>
<th>training</th>
<th>TE (CV)</th>
<th>testing</th>
<th>TE (CV)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1997</td>
<td>0 (8.98)</td>
<td>1998</td>
<td>0 (9.02)</td>
</tr>
<tr>
<td>1997-1998</td>
<td>0 (8.99)</td>
<td>1999</td>
<td>0 (10.03)</td>
</tr>
<tr>
<td>1997-1999</td>
<td>0 (9.37)</td>
<td>2000</td>
<td>0 (6.89)</td>
</tr>
<tr>
<td>1997-2000</td>
<td>0 (8.57)</td>
<td>2001</td>
<td>5.43 (5.86)</td>
</tr>
<tr>
<td>1997-2001</td>
<td>0 (4.55)</td>
<td>2002</td>
<td>4.68 (4.61)</td>
</tr>
</tbody>
</table>

Table 5: Percentage of Training Error (TE) and Cross-Validation (5-fold CV)
Current findings

- SVM with optimal parameter is more robust to the imbalanced data set
- Evolutionary feature selection (using Genetic Algorithm) could find global solution of SVM parameter optimization
- More investigation to variable selection
Variable Selection and Optimization in Default Prediction

Dedy Dwi Prastyo
Wolfgang Karl Härdle

Ladislaus von Bortkiewicz Chair of Statistics
C.A.S.E. – Center for Applied Statistics and Economics
Humboldt-Universität zu Berlin
http://lvb.wiwi.hu-berlin.de
http://www.case.hu-berlin.de
References

Chen, S., Härdle, W. and Moro, R.
Estimation of Default Probabilities with Support Vector Machines
Quantitative Finance, 2011, 11, 135 - 154

Efron, B., Hastie, T., Johnstone, I. and Tibshirani, R.
Least angle regression

Friedman, J., Hastie, T., Hoefling, H. and Tibshirani, R.
Pathwise coordinate optimization
Friedman, J., Hastie, T. and Tibshirani, R.
Regularization path for generalized linear models via coordinate descent

Härdle, Lee, Y-J., Schäfer, D. and Yeh, Y-R.
Variable Selection and Oversampling in the Use of Smooth Support Vector Machine for Predicting the Default Risk of Companies
Journal of Forecasting, 2009, 28, 512 - 534

Holland, J.H.
Adaptation in Natural and Artificial Systems
University of Michigan Press, 1975
References

- Ka ratzoglou, A. and Meyer, D.
  Support Vector Machines in R

- Rosset, S. and Zhu, J.
  Piecewise linear regularized solution path

- Tibshirani, R.
  Regression shrinkage and selection via Lasso
References

Tseng, P.
Convergence of a block coordinate descent method for nondifferentiable minimization

Van der Kooij, A.
Prediction accuracy and stability of regression with optimal scaling transformation
Ph.D. thesis, Department of data theory, University of Leiden, 2001
Linearly Separable Case

Figure 9: Separating hyperplane and its margin in linearly separable case
Choose \( f \in \mathcal{F} \) such that margin \((d_- + d_+)\) is maximal

No error separation, if all \( i = 1, 2, ..., n \) satisfy

\[
\begin{align*}
    x_i^\top w + b & \geq +1 & \text{for } y_i = +1 \\
    x_i^\top w + b & \leq -1 & \text{for } y_i = -1
\end{align*}
\]

Both constraints are combined into

\[
y_i(x_i^\top w + b) - 1 \geq 0 \quad i = 1, 2, ..., n
\]
Distance between margins and the separating hyperplane is
\[ d_+ = d_- = \frac{1}{\|w\|} \]
Maximize the margin, \[ d_+ + d_- = \frac{2}{\|w\|} \], could be attained by minimizing \( \|w\| \) or \( \|w\|^2 \)
Lagrangian for the primal problem
\[
L_P(w, b) = \frac{1}{2}\|w\|^2 - \sum_{i=1}^{n} \alpha_i \{y_i(x_i^\top w + b) - 1\}
\]
Karush-Kuhn-Tucker (KKT) first order optimality conditions

\[
\frac{\partial L_P}{\partial w_k} = 0 : \quad w_k - \sum_{i=1}^{n} \alpha_i y_i x_{ik} = 0 \quad k = 1, \ldots, d
\]

\[
\frac{\partial L_P}{\partial b} = 0 : \quad \sum_{i=1}^{n} \alpha_i y_i = 0
\]

\[
y_i (x_i^\top w + b) - 1 \geq 0 \quad i = 1, \ldots, n
\]

\[
\alpha_i \geq 0
\]

\[
\alpha_i \{ y_i (x_i^\top w + b) - 1 \} = 0
\]
Solution \( w = \sum_{i=1}^{n} \alpha_i y_i x_i \), therefore

\[
\frac{1}{2} \|w\|^2 = \frac{1}{2} \sum_{i=1}^{n} \sum_{j=1}^{n} \alpha_i \alpha_j y_i y_j x_i^\top x_j
\]

\[= - \sum_{i=1}^{n} \alpha_i \{y_i (x_i^\top w + b) - 1\} = - \sum_{i=1}^{n} \alpha_i y_i x_i^\top \sum_{j=1}^{n} \alpha_j y_j x_j + \sum_{i=1}^{n} \alpha_i
\]

\[= - \sum_{i=1}^{n} \sum_{j=1}^{n} \alpha_i \alpha_j y_i y_j x_i^\top x_j + \sum_{i=1}^{n} \alpha_i
\]

Lagrangian for the dual problem

\[
L_D(\alpha) = \sum_{i=1}^{n} \alpha_i - \frac{1}{2} \sum_{i=1}^{n} \sum_{j=1}^{n} \alpha_i \alpha_j y_i y_j x_i^\top x_j
\]
Primal and dual problems

\[
\min_{w, b} \quad L_P(w, b)
\]

\[
\max_{\alpha} \quad L_D(\alpha) \quad \text{s.t.} \quad \alpha_i \geq 0, \quad \sum_{i=1}^{n} \alpha_i y_i = 0
\]

Optimization problem is convex, therefore the dual and primal formulations give the same solution

Support vector, a point \( i \) for which \( y_i(x_i^\top w + b) = 1 \) holds
Figure 10: Hyperplane and its margin in linearly non-separable case
Slack variables $\xi_i$ represent the violation from strict separation

\[ x_i^T w + b \geq 1 - \xi_i \quad \text{for} \quad y_i = 1, \]
\[ x_i^T w + b \leq -1 + \xi_i \quad \text{for} \quad y_i = -1, \]
\[ \xi_i \geq 0 \]

constraints are combined into

\[ y_i (x_i^T w + b) \geq 1 - \xi_i \quad \text{and} \quad \xi_i \geq 0 \]

If $\xi_i > 0$, the objective function is

\[ \frac{1}{2} \|w\|^2 + C \sum_{i=1}^{n} \xi_i \]
Lagrange function for the primal problem

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

where \( \alpha_i \geq 0 \) and \( \mu_i \geq 0 \) are Lagrange multipliers

Primal problem

\[ \min_{w, b, \xi} L_P (w, b, \xi) \]
First order conditions

\[
\frac{\partial L_P}{\partial w_k} = 0 : \quad w_k - \sum_{i=1}^{n} \alpha_i y_i x_{ik} = 0
\]

\[
\frac{\partial L_P}{\partial b} = 0 : \quad \sum_{i=1}^{n} \alpha_i y_i = 0
\]

\[
\frac{\partial L_P}{\partial \xi_i} = 0 : \quad C - \alpha_i - \mu_i = 0
\]

s.t. \( \alpha_i \geq 0, \quad \mu_i \geq 0, \quad \mu_i \xi_i = 0 \)

\[
\alpha_i \{ y_i (x_i^\top w + b) - 1 + \xi_i \} = 0
\]
Note that $\sum_{i=1}^{n} \alpha_i y_i b = 0$. Translate primal problem into

$$L_D (\alpha) = \sum_{i=1}^{n} \alpha_i - \frac{1}{2} \sum_{i=1}^{n} \sum_{j=1}^{n} \alpha_i \alpha_j y_i y_j x_i^\top x_j + \sum_{i=1}^{n} \xi_i (C - \alpha_i - \mu_i)$$

Last term is 0, therefore the dual problem is

$$\max_{\alpha} L_D (\alpha) = \max_{\alpha} \left\{ \sum_{i=1}^{n} \alpha_i - \frac{1}{2} \sum_{i=1}^{n} \sum_{j=1}^{n} \alpha_i \alpha_j y_i y_j x_i^\top x_j \right\},$$

s.t. $0 \leq \alpha_i \leq C$, $\sum_{i=1}^{n} \alpha_i y_i = 0$
GA – Initialization

Figure 11: GA at first generation
GA – Convergencia

Figure 12: Solutions at 1st generation (left) and rth generation (right)
GA – Decoding

\[ \theta = \theta_{\text{lower}} + (\theta_{\text{upper}} - \theta_{\text{lower}}) \sum_{i=0}^{l-1} a_i 2^i \]

where \( \theta \) is solution (i.e. parameter), \( a \) is allele
GA – Fitness evaluation

- Calculate $f(\theta_i)$, $i = 1, \ldots, \text{popsize}$
- Evaluate fitness, $f_{dp}(\theta_i)$
  \[ f_{dp}(\theta_i) = \text{AR, AUC, accuracy, specificity, sensitivity} \]
- Relative fitness, $p_i = \frac{f_{dp}(\theta_i)}{\sum_{k=i}^{\text{popsize}} f_{dp}(\theta^i)}$

Figure 14: Proportion to be chosen in the next iteration (generation)
GA – Roulette wheel

- $\text{rand} \sim \text{U}(0, 1)$
- Select $i^{th}$ chromosome if $\sum_{i=1}^{k} p_i < \text{rand} < \sum_{i=1}^{k+1} p_i$
- Repeat $\text{popsiz}e$ times to get $\text{popsiz}e$ new chromosomes
GA – Crossover

Figure 15: Crossover in nature

Figure 16: Randomly chosen one-point crossover (top) and two-points crossover (bottom)
GA – Reproductive operator

Figure 17: One-point crossover (top) and bit-flip mutation (bottom)
GA – Elitism

- Best solution in each iteration is maintained in another memory place
- New population replaces the old one, check whether best solution is in the population
  If not, replace any one in the population with best solution
## Nature to Computer Mapping

<table>
<thead>
<tr>
<th>Nature</th>
<th>GA-SVM</th>
</tr>
</thead>
<tbody>
<tr>
<td>Population</td>
<td>Set of parameter</td>
</tr>
<tr>
<td>Individual (phenotype)</td>
<td>Parameters</td>
</tr>
<tr>
<td>Fitness</td>
<td>Discriminatory power</td>
</tr>
<tr>
<td>Chromosome (genotype)</td>
<td>Encoding of parameter</td>
</tr>
<tr>
<td>Gene</td>
<td>Binary encoding</td>
</tr>
<tr>
<td>Reproduction</td>
<td>Crossover</td>
</tr>
<tr>
<td>Generation</td>
<td>Iteration</td>
</tr>
</tbody>
</table>

Table 6: Nature to GA-SVM mapping
Examples

- Small sample: 100 solvent and insolvent companies
- Credit reform data
- X3 – Operating Income / Total Asset
- X24 – Account Payable / Total Asset
Figure 18: SVM plot, $C = 1$ and $\sigma = 1/2$, misclass. rate 0.19 (left) and GA-SVM, $C = 14.86$ and $\sigma = 1/121.61$, misclass. rate 0 (right).
Figure 19: GA-SVM ($C = 187.93$ and $\sigma = 1/195.16$) plot of training data, misclass. rate 2.38%, and testing data, misclass. rate 1.37%.
**FR: Profitability**

<table>
<thead>
<tr>
<th>Ratio No.</th>
<th>Definition</th>
<th>Ratio</th>
</tr>
</thead>
<tbody>
<tr>
<td>x1</td>
<td>NI/TA</td>
<td>Return on assets (ROA)</td>
</tr>
<tr>
<td>x2</td>
<td>NI/Sales</td>
<td>Net profit margin</td>
</tr>
<tr>
<td>x3</td>
<td>OI/TA</td>
<td>Operating Income/Total assets</td>
</tr>
<tr>
<td>x4</td>
<td>OI/Sales</td>
<td>Operating profit margin</td>
</tr>
<tr>
<td>x5</td>
<td>EBIT/TA</td>
<td>EBIT/Total assets</td>
</tr>
<tr>
<td>x6</td>
<td>(EBIT+AD)/TA</td>
<td>EBITDA</td>
</tr>
<tr>
<td>x7</td>
<td>EBIT/Sales</td>
<td>EBIT/Sales</td>
</tr>
</tbody>
</table>

Table 7: Definitions of financial ratios.
### FR: Leverage

<table>
<thead>
<tr>
<th>Ratio No.</th>
<th>Definition</th>
<th>Ratio</th>
</tr>
</thead>
<tbody>
<tr>
<td>x8</td>
<td>Equity/TA</td>
<td>Own funds ratio (simple)</td>
</tr>
<tr>
<td>x9</td>
<td>(Equity-ITGA)/(TA-ITGA-Cash-LB)</td>
<td>Own funds ratio (adjusted)</td>
</tr>
<tr>
<td>x10</td>
<td>CL/TA</td>
<td>Current liabilities/Total assets</td>
</tr>
<tr>
<td>x11</td>
<td>(CL-Cash)/TA</td>
<td>Net indebtedness</td>
</tr>
<tr>
<td>x12</td>
<td>TL/TA</td>
<td>Total liabilities/Total assets</td>
</tr>
<tr>
<td>x13</td>
<td>Debt/TA</td>
<td>Debt ratio</td>
</tr>
<tr>
<td>x14</td>
<td>EBIT/Interest expenses</td>
<td>Interest coverage ratio</td>
</tr>
</tbody>
</table>

Table 8: Definitions of financial ratios.
## FR: Liquidity

<table>
<thead>
<tr>
<th>Ratio No.</th>
<th>Definition</th>
<th>Ratio</th>
</tr>
</thead>
<tbody>
<tr>
<td>x15</td>
<td>Cash/TA</td>
<td>Cash/Total assets</td>
</tr>
<tr>
<td>x16</td>
<td>Cash/CL</td>
<td>Cash ratio</td>
</tr>
<tr>
<td>x17</td>
<td>QA/CL</td>
<td>Quick ratio</td>
</tr>
<tr>
<td>x18</td>
<td>CA/CL</td>
<td>Current ratio</td>
</tr>
<tr>
<td>x19</td>
<td>WC/TA</td>
<td>Working Capital</td>
</tr>
<tr>
<td>x20</td>
<td>CL/TL</td>
<td>Current liabilities/Total liabilities</td>
</tr>
</tbody>
</table>

Table 9: Definitions of financial ratios.
## FR: Activity

<table>
<thead>
<tr>
<th>Ratio No.</th>
<th>Definition</th>
<th>Ratio</th>
</tr>
</thead>
<tbody>
<tr>
<td>x21</td>
<td>TA/Sales</td>
<td>Asset turnover</td>
</tr>
<tr>
<td>x22</td>
<td>INV/Sales</td>
<td>Inventory turnover</td>
</tr>
<tr>
<td>x23</td>
<td>AR/Sales</td>
<td>Account receivable turnover</td>
</tr>
<tr>
<td>x24</td>
<td>AP/Sales</td>
<td>Account payable turnover</td>
</tr>
<tr>
<td>x25</td>
<td>Log(TA)</td>
<td>Log(Total assets)</td>
</tr>
</tbody>
</table>

Table 10: Definitions of financial ratios.
### Ratio No. | Definition
--- | ---
$x26$ | increase (decrease) in inventories / inventories
$x27$ | increase (decrease) in liabilities / total liabilities
$x28$ | increase (decrease) in cash flows / cash and cash equivalent

Table 11: Definitions of financial ratios.
Findings

- Härdle et al. (2009): Smooth SVM – overall mean of correct predictions ranging from 70% to 78% (misclassification: 22% to 30%)

- Chen, Härdle and Moro (2011):
  - Most of the models tested, AR in 43.50% and 60.51%
  - SVM (grid search optimization): percentage of correctly classified out-of-sample 71.85%
  - Logit model: percentage of correctly classified out-of-sample 67.24%
Findings

Zang and Härdle (2010):

<table>
<thead>
<tr>
<th>Performance measure</th>
<th>Logit (%)</th>
<th>CART (%)</th>
<th>BACT (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Overall misclassification rate</td>
<td>30.2</td>
<td>33.8</td>
<td>26.6</td>
</tr>
<tr>
<td>Type I misclassification rate</td>
<td>28.3</td>
<td>27.2</td>
<td>27.6</td>
</tr>
<tr>
<td>Type II misclassification rate</td>
<td>30.3</td>
<td>34.3</td>
<td>26.5</td>
</tr>
<tr>
<td>AR</td>
<td>52.1</td>
<td>58.7</td>
<td>60.4</td>
</tr>
</tbody>
</table>

Table 12: Average value (of bootstrap) of performance measures