Clustering Large Credit Client Data Sets for Classification with SVM

Ralf Stecking
University of Oldenburg
Department of Economics

Klaus B. Schebesch
University "Vasile Goldiş" Arad
Faculty of Economics

University of Edinburgh
Credit Scoring and Credit Control XI Conference

26.08.2009

Overview

- Motivation
- Kernels and clustering
- Preliminary evaluation of cluster based SVM
- Multiple validation of cluster based SVM
- Credit scoring and data clustering
- Symbolic representation of credit client clusters
- Symbolic SVM model building and evaluation
- Conclusions and outlook

Motivation

- In past work we used a medium sized empirical credit scoring data set with $N = 658$ credit clients all having $m = 40$ input features in order to analyze different aspects of model building with regard to out of sample classification performance.
- As base model we used SVM and other statistical learning methods like LDA, CART, and LogReg. Thereupon model combinations of outputs of the base models were also investigated.
- Gaining access to a $N \approx 140.000$ data set with $m = 23$ features per credit client and with extremly asymmetric class distributions, precludes case-by-case training of models like SVM.
- Further increasing N and "fusing" credit client information from different sources will worsen the situation.

Kernels and relations between pairs of cases (1)

- A training set $\{y_i|x_i\}_{i=1,...,N}$ may contain labeled credit clients (e.g. $y_i \in \{-1, 1\}$) or unlabeled ones ($y_i = 0$, all i, say).
- A kernel function $k_{s_{ij}}(x_i, x_j) \geq 0$ describes a metric relation (inverted distance, etc.) between any two training feature vectors $x_i, x_j \in \{1, ..., N\}$.
- The implied numerical matrix K_{ij} is usually meant to be symmetric.
- Individualized parameters for pairs of clients ij can impose conditions which may act
  - classwise (e.g. correct for asymmetic costs)
  - casewise (e.g. correct for case importance) or
  - interaction-wise (i.e. 2-interactions).

Kernels and relations between pairs of cases (2)

An instantiation of $k_{ij}(x_i, x_j)$ would be the adaptation of the very powerful RBF kernel for nonlinear SVM:

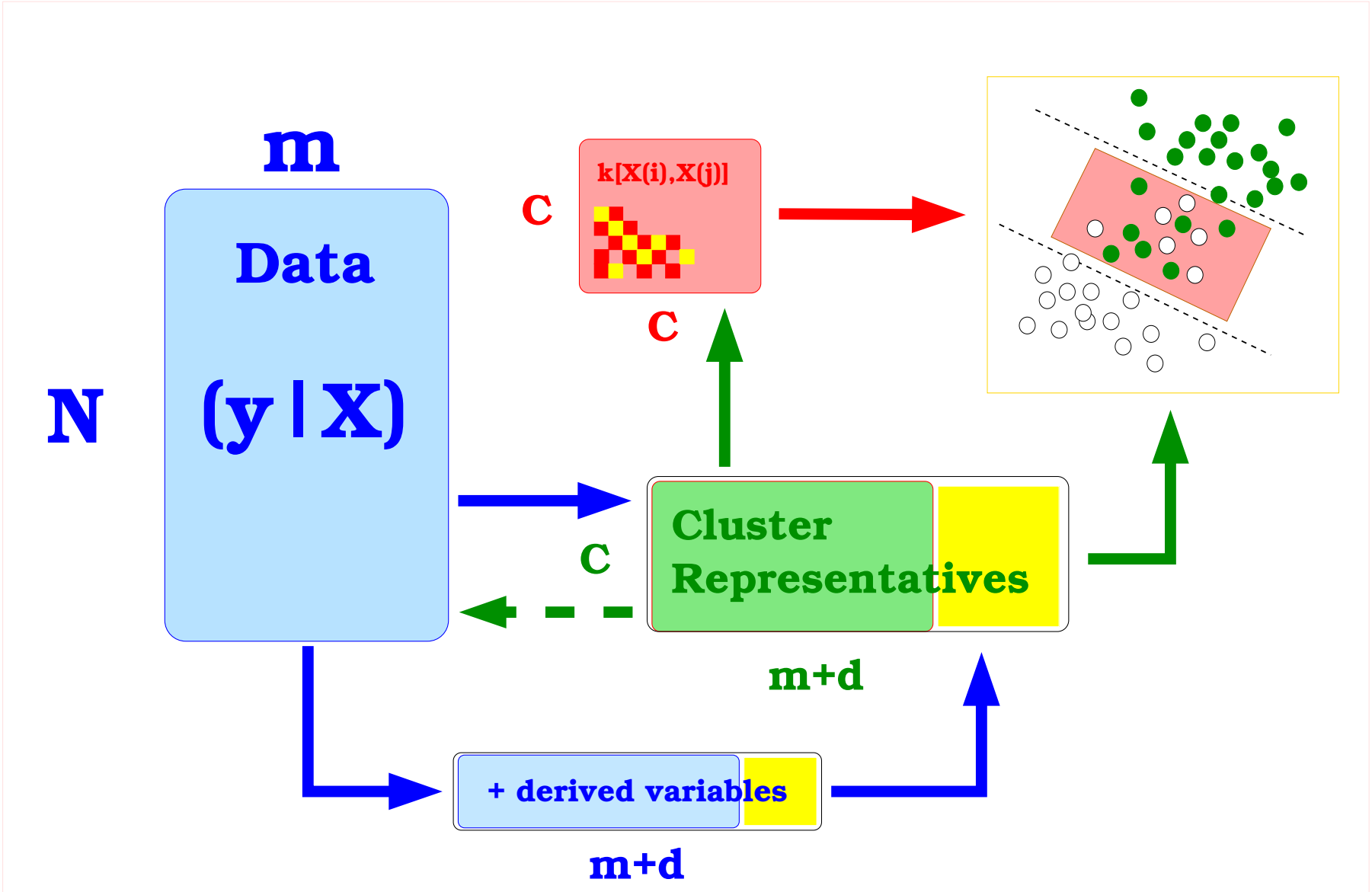
$$r_{ij} \exp(-s_{ij}||x_i - x_j||^2), \quad \text{with}$$
$$r_{ij} \in \{0, 1\} \quad \text{a grouping relation, and}$$
$$s_{ij} \geq 0 \quad \text{the interaction sensitivity.}$$

By identically permuting the index sets of i and j, i.e. via i_k and j_k, the matrix $(K_{i_k j_k})$ resulting from the kernel may be block-diagonal, indicating a cluster structure, which in turn may or may not be dependent of the labeling $\{y_{i_k}\}$.
The most popular RBF SVM is simply using $r_{ij} = 1$ for all i, j and $s_{ij} = s > 0$, a constant, for all i, j.

Kernels and relations between pairs of cases (3)

- Being able to set (at least part of) the r_{ij} and s_{ij} parameters would convey domain knowledge into the problem formulation. Do we have such knowledge ?
- Treating r_{ij} and s_{ij} as "slow" variables which are to be optimized along with "fast" variables, i.e. the SVM duals and slacks, is hardly tractable (for empirically reasonale N).
- In order to approxiamte this task, we split the associated simulataneous problem into two consecutive tasks which are routinely tractable:
  - Cluster the data set by a fast standard method (with and without label prepartitioning)
  - Apply the SVM kernel to the resulting cluster prototypes.

What we do for this presentation



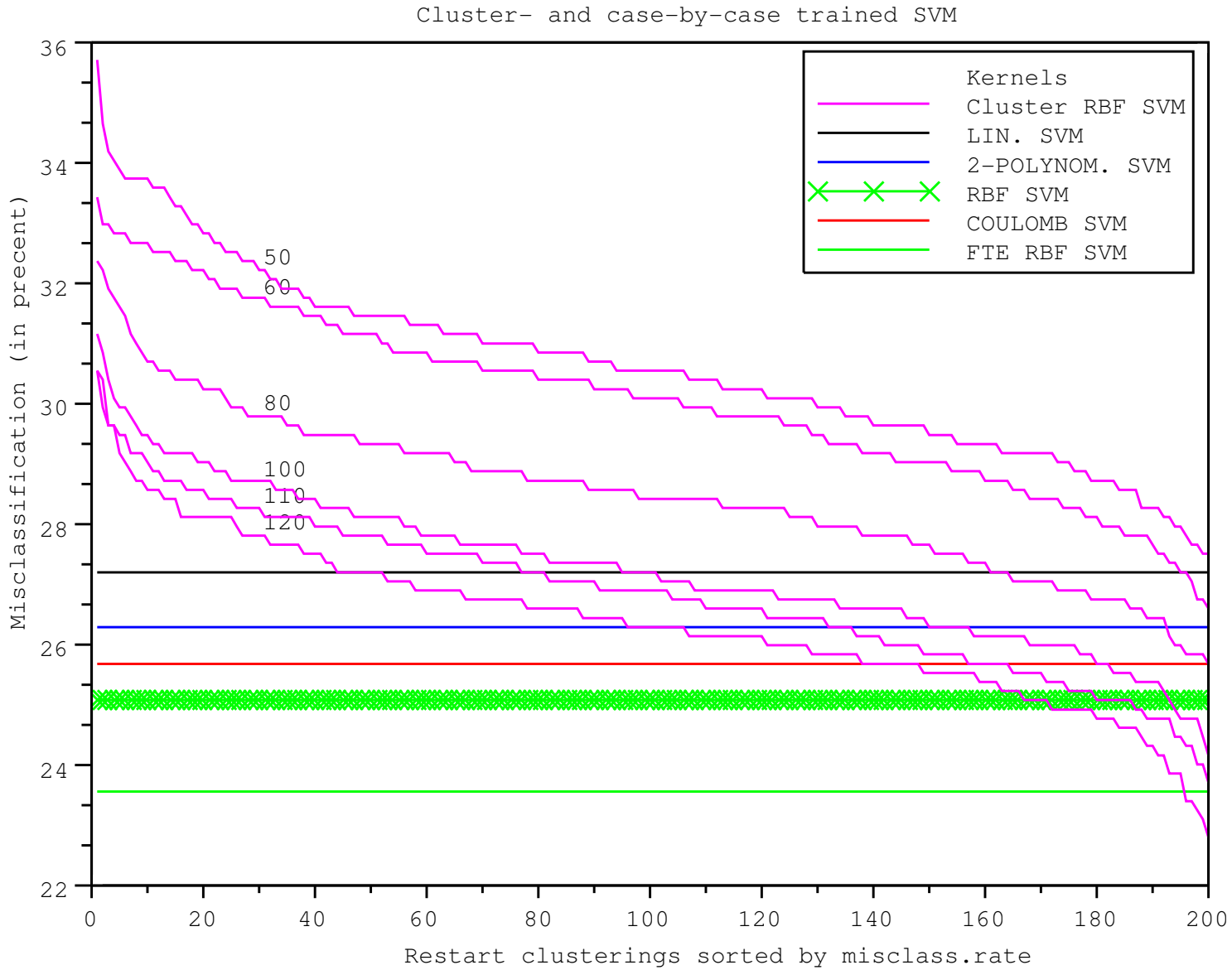
Some issues concerning cluster formation

Is there any cluster structure in the data ?
A clustering algorithm will issue "clusters" for very $1 \leq c \leq N$!
Suppose there is some (empirical) clustering, are the cluster shapes compact (spheroidal) or elongated or of mixed shape in high dimensions and are they well separated ?
The clustering method can be one of the following

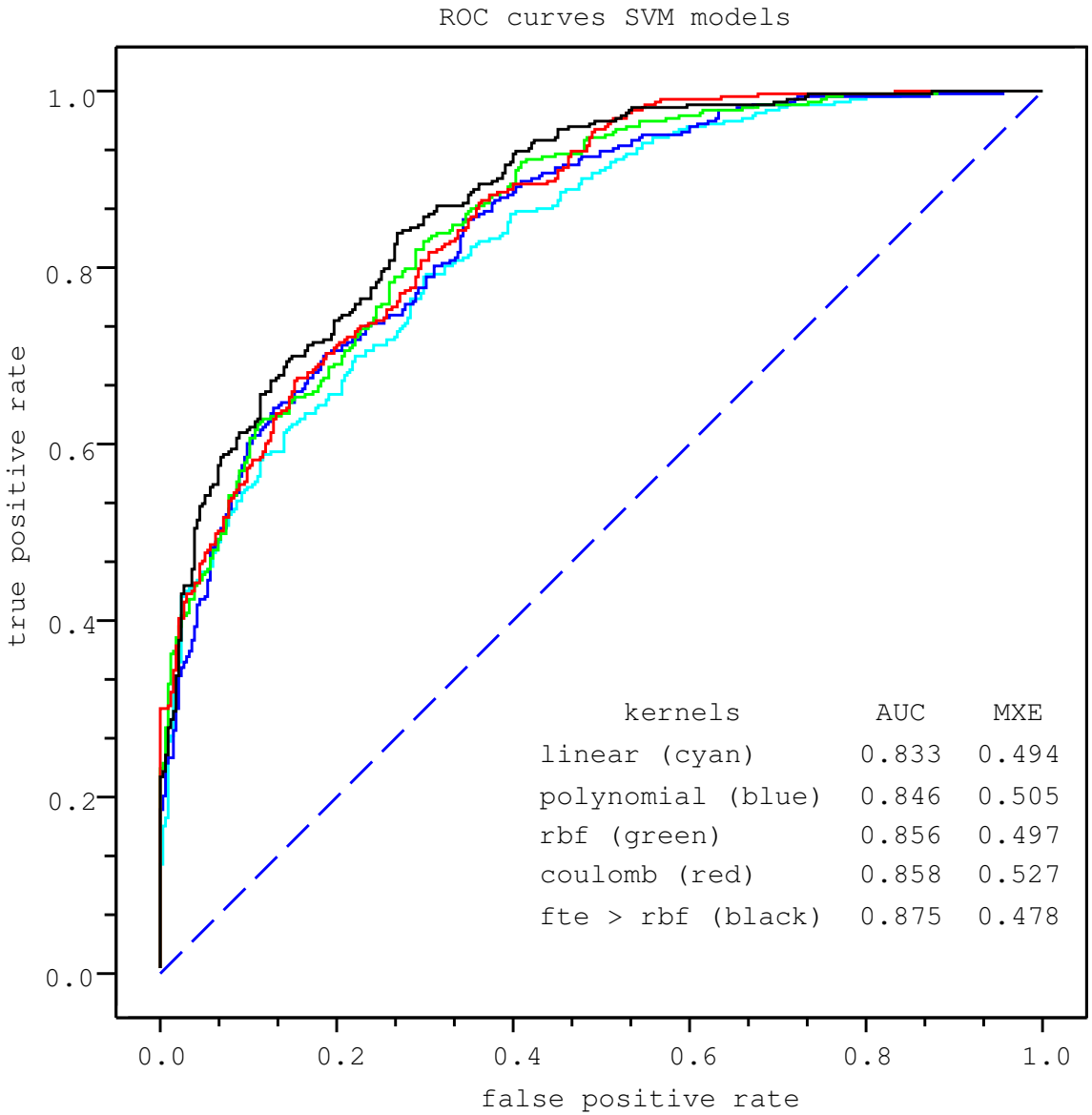
- completely unsupervised
- ...
- constrained to some degree ("soft" penalty terms, using balancing and correlations etc. arguments, of "must-be" or "cannot-be" type, ... , can cluster members of predefined classes only, ...)

How to cluster labeled credit client data ?
Cluster the entire training data or cluster the members of predefined classes only ?

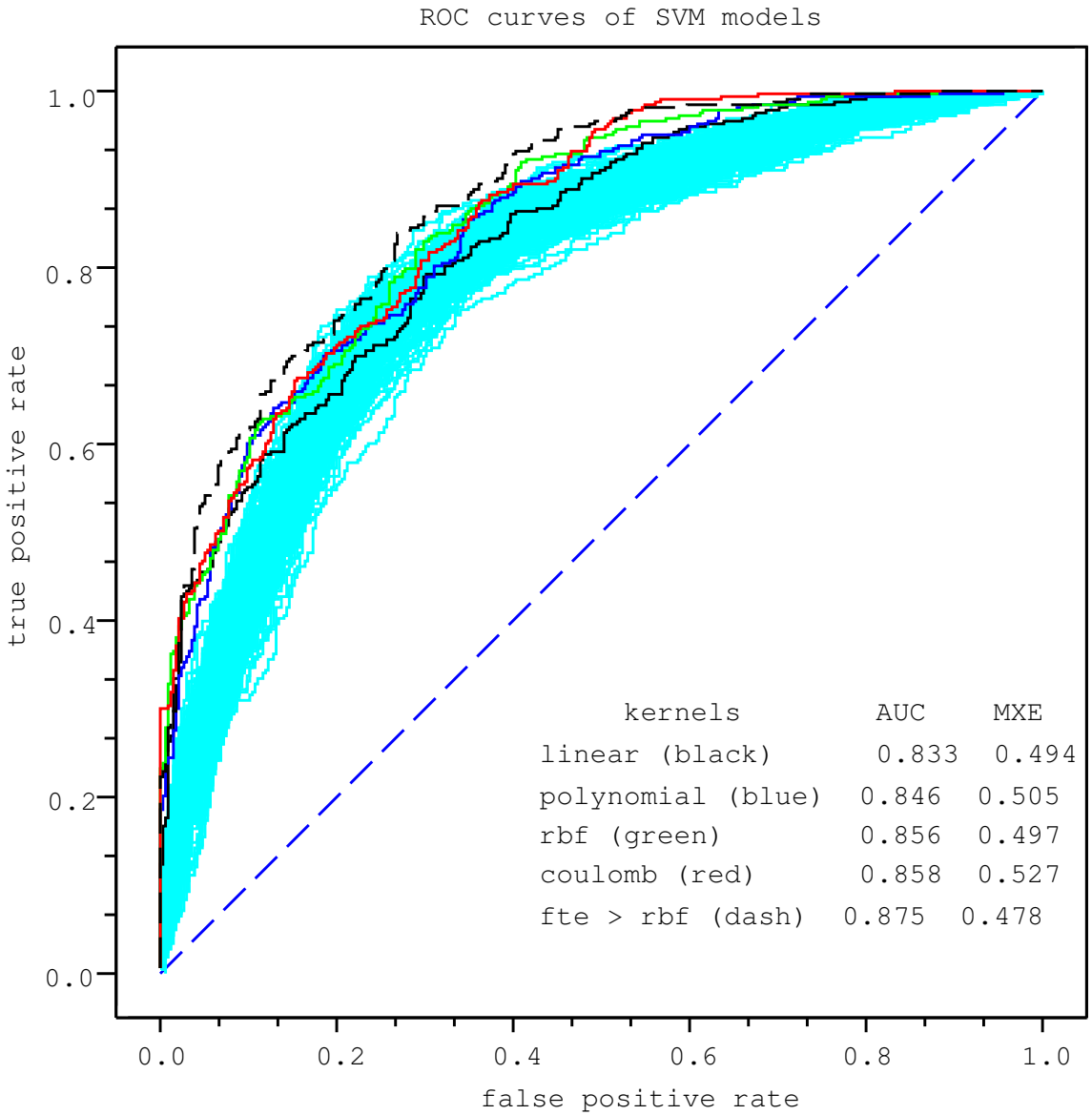
Different models trained on cluster prototypes



ROC curves of SVM with different kernels ...



... and ROC curves of cluster based RBF SVM

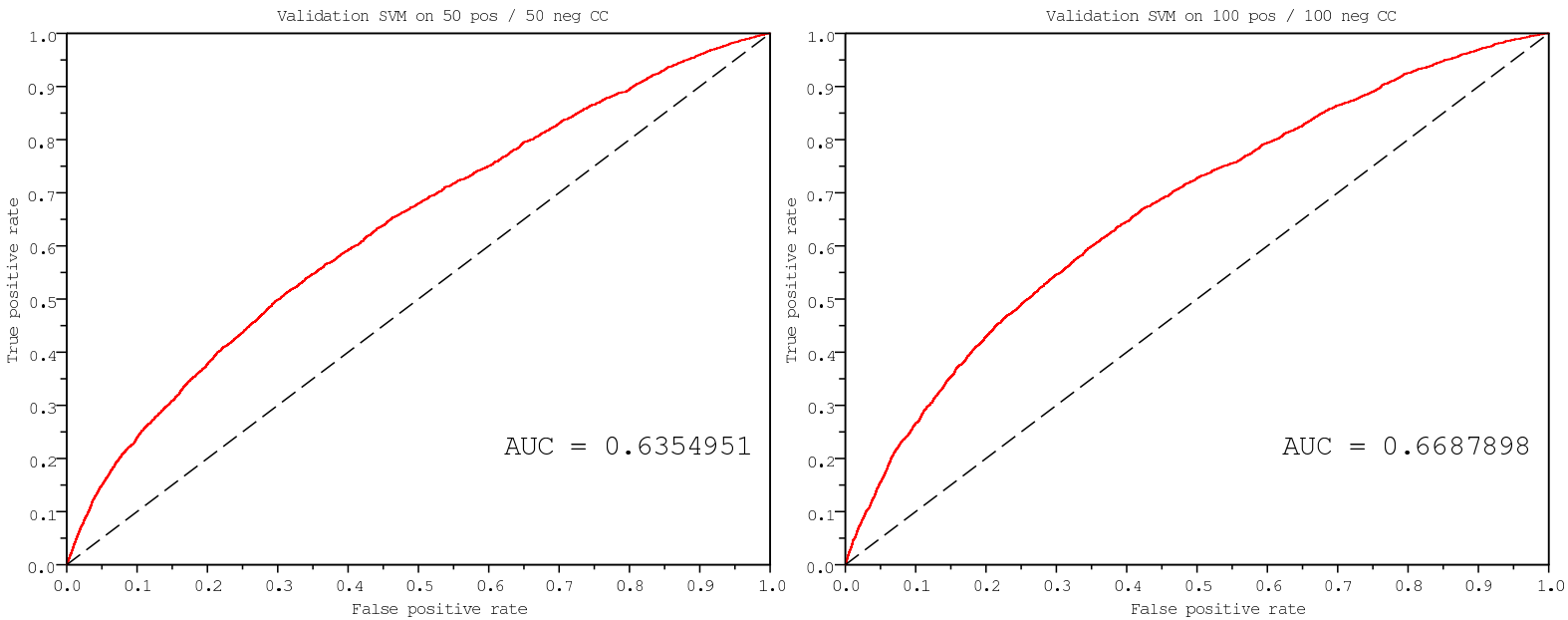


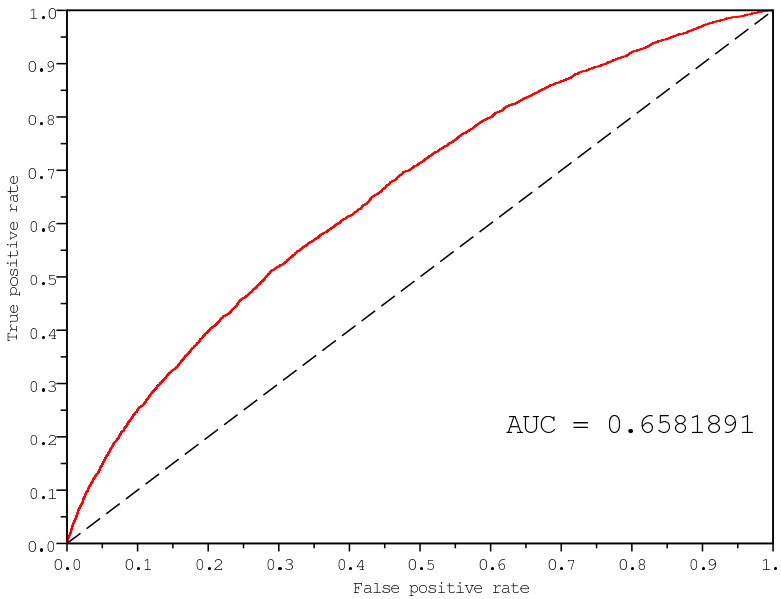
Validation of cluster based SVM on a large credit client data set

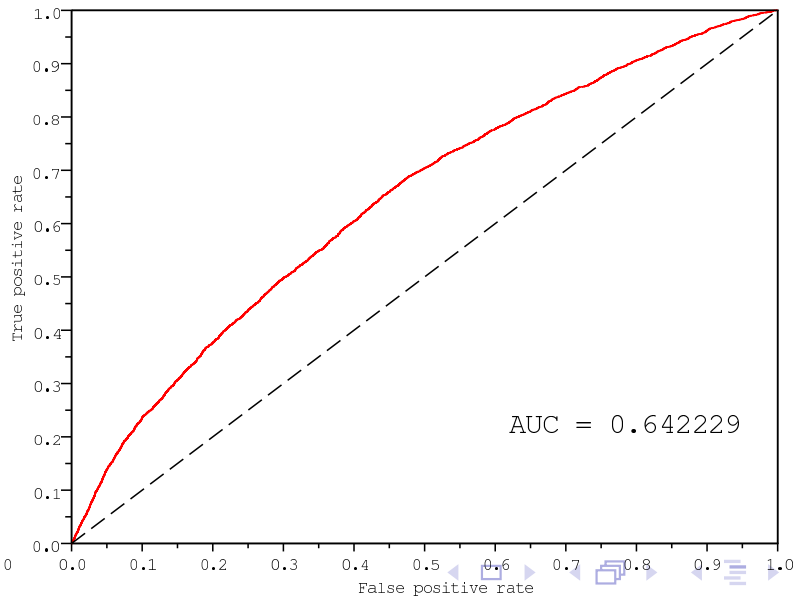
It is not sufficient to validate the SVM models trained on a given set of cluster representatives! Improved validation includes clustering, the outcome of which may be different when using different holdout sets:

1. Step A: Divide the training set into positive and negative cases $T = P \cup N$.
2. Step B: Subdivide both P and N of a large training set (of > 100000 cases, say) into n (approx. equally sized) non-overlapping segments $[P_1, P_2, ..., P_i, ..., P_n]$ and $[N_1, N_2, ..., N_i, ..., N_n]$ with the smallest segment containing at least 30 cases, say.
3. Step C i: cluster both sets $[P_1, P_2, ..., P_{i-1}, P_{i+1}..., P_n]$ and $[N_1, N_2, ..., N_{i-1}, N_{i+1}..., N_n]$ obtaining $2c$ cluster representatives. Train a SVM on these labeled $2c$ points.
4. Step D i: validate the ith SVM just on the segment $[P_i, N_i]$.

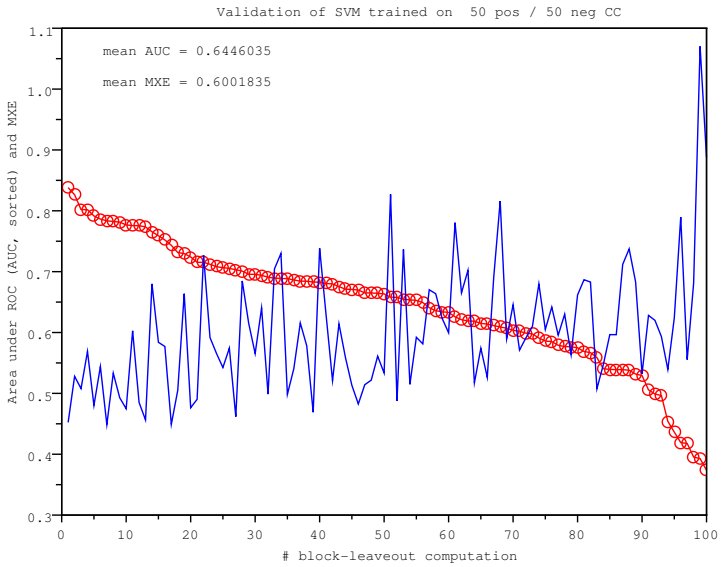
ROC curves for SVM 50,100,200 and 400 clusters

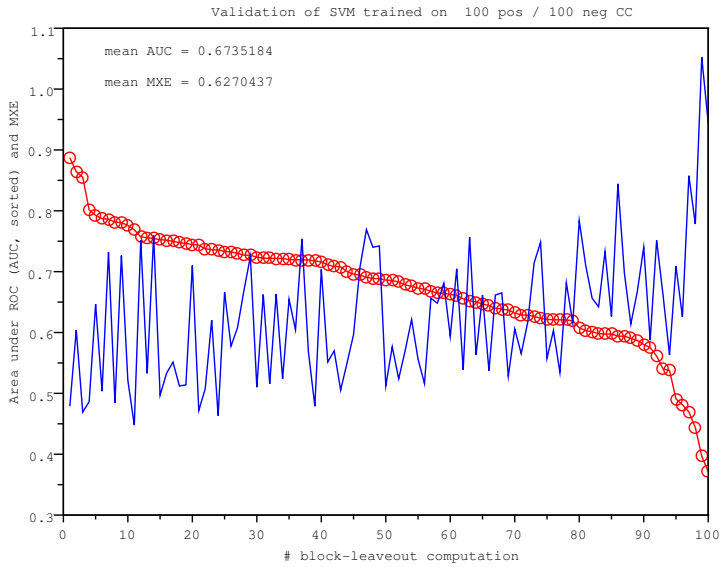


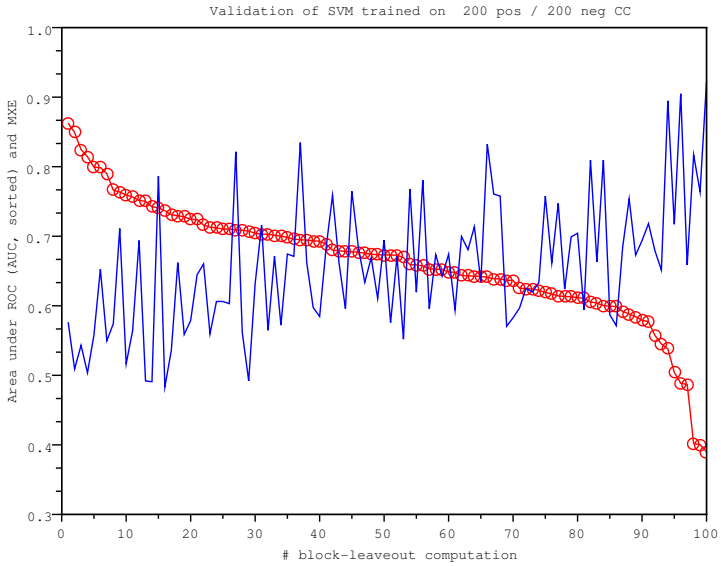


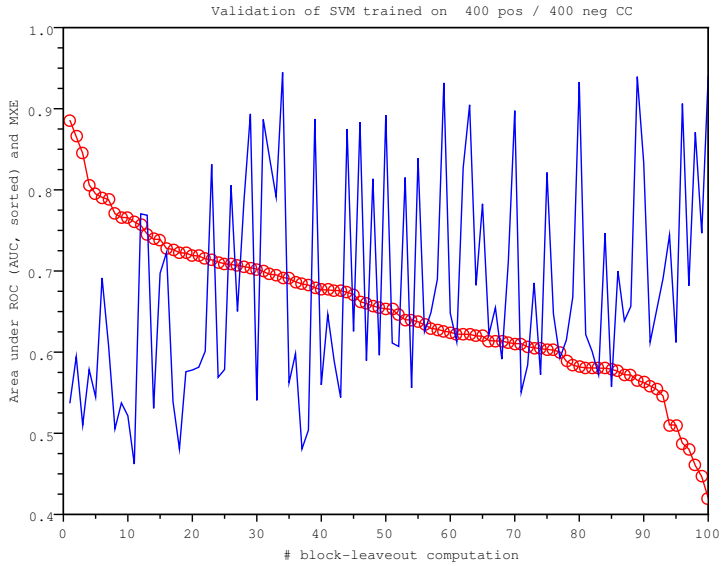


Sorted AUC for SVM 50,100,200 and 400 clusters

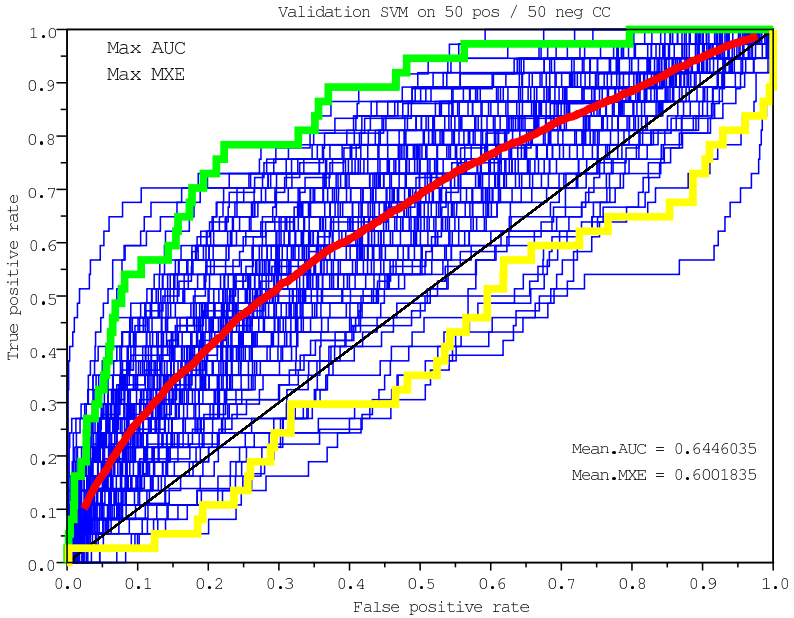


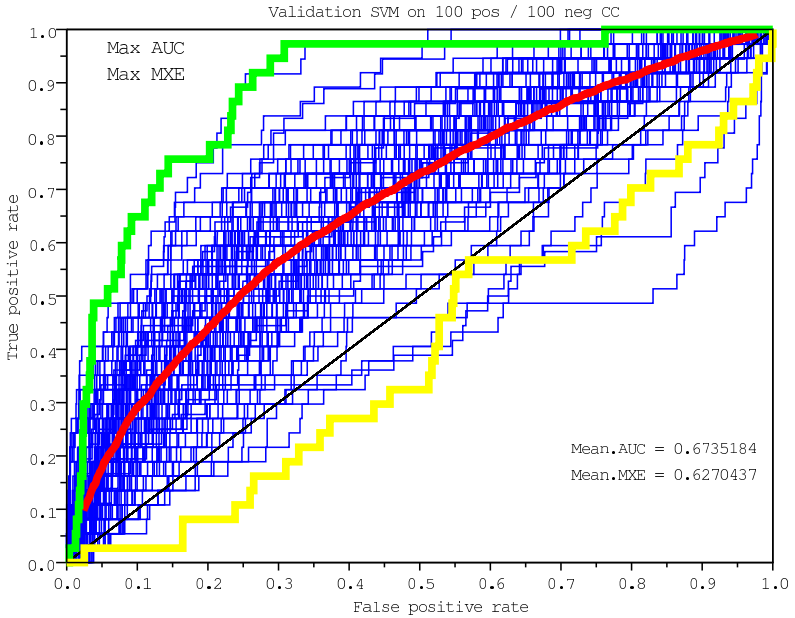


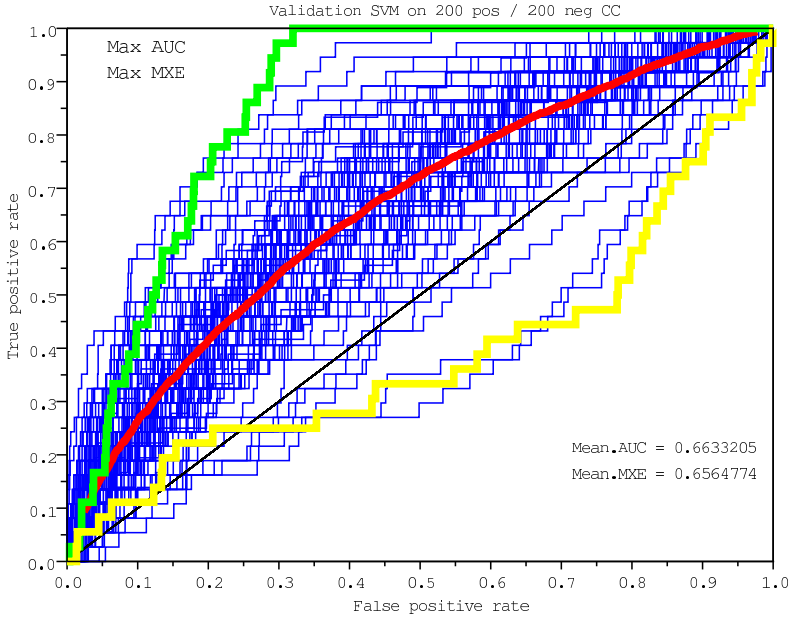


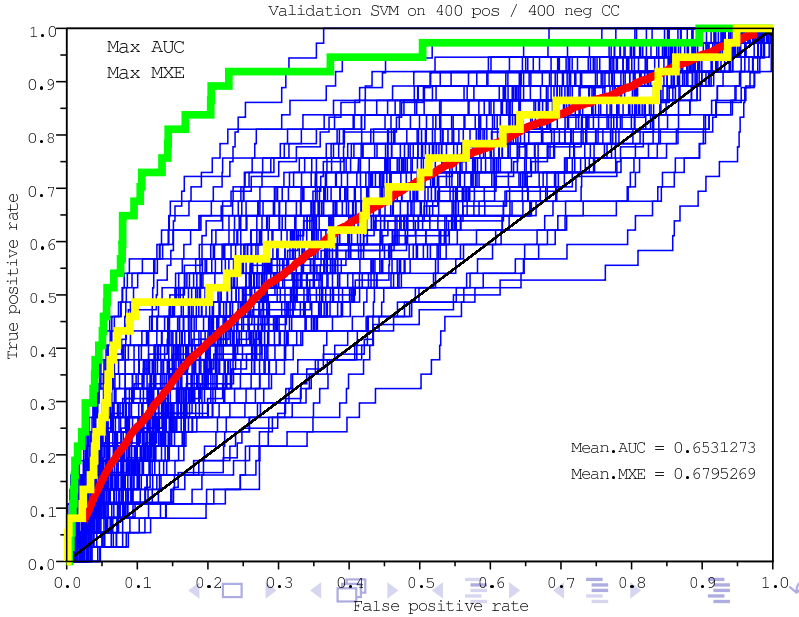


Sliced ROC for SVM 50,100,200 and 400 clusters

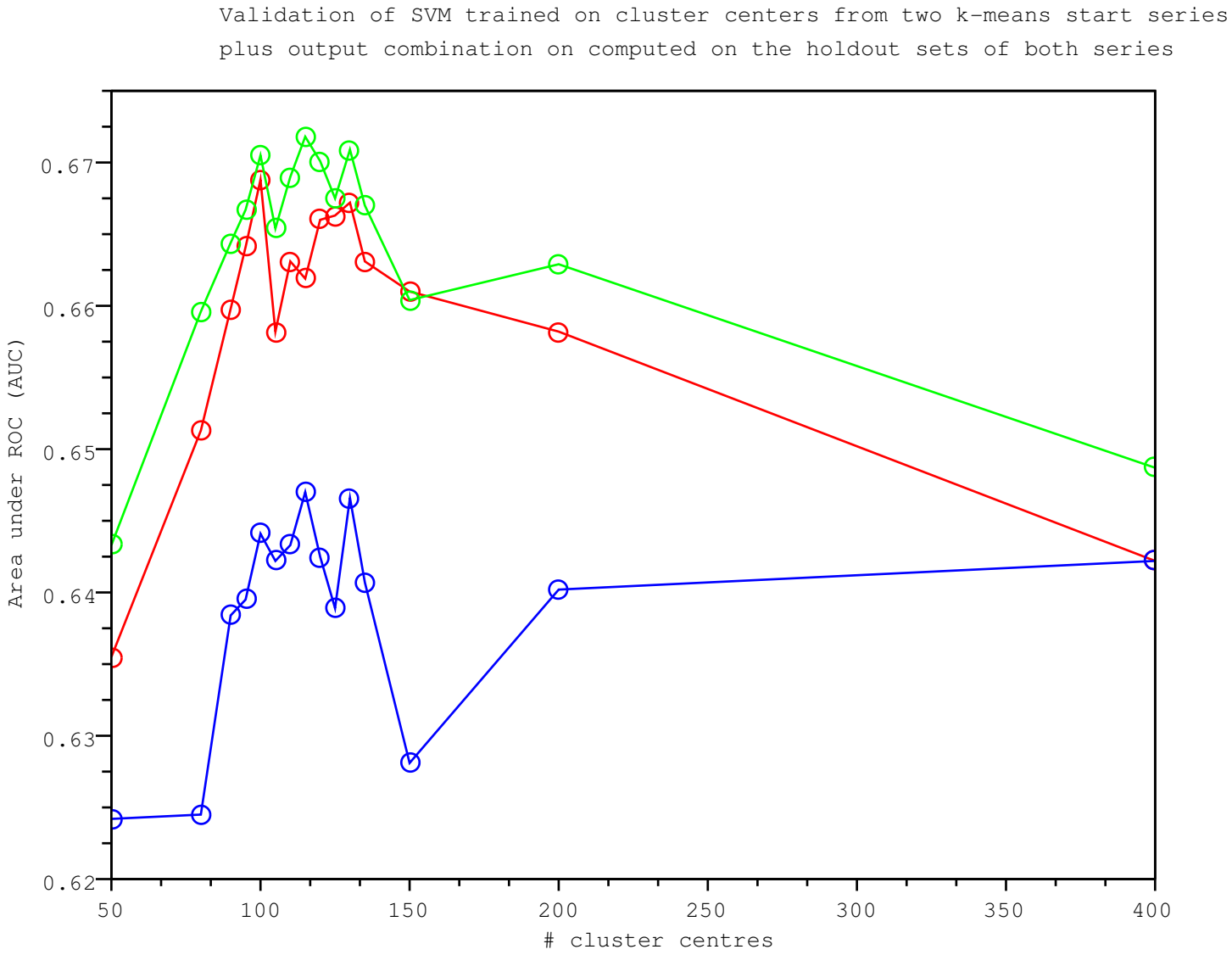




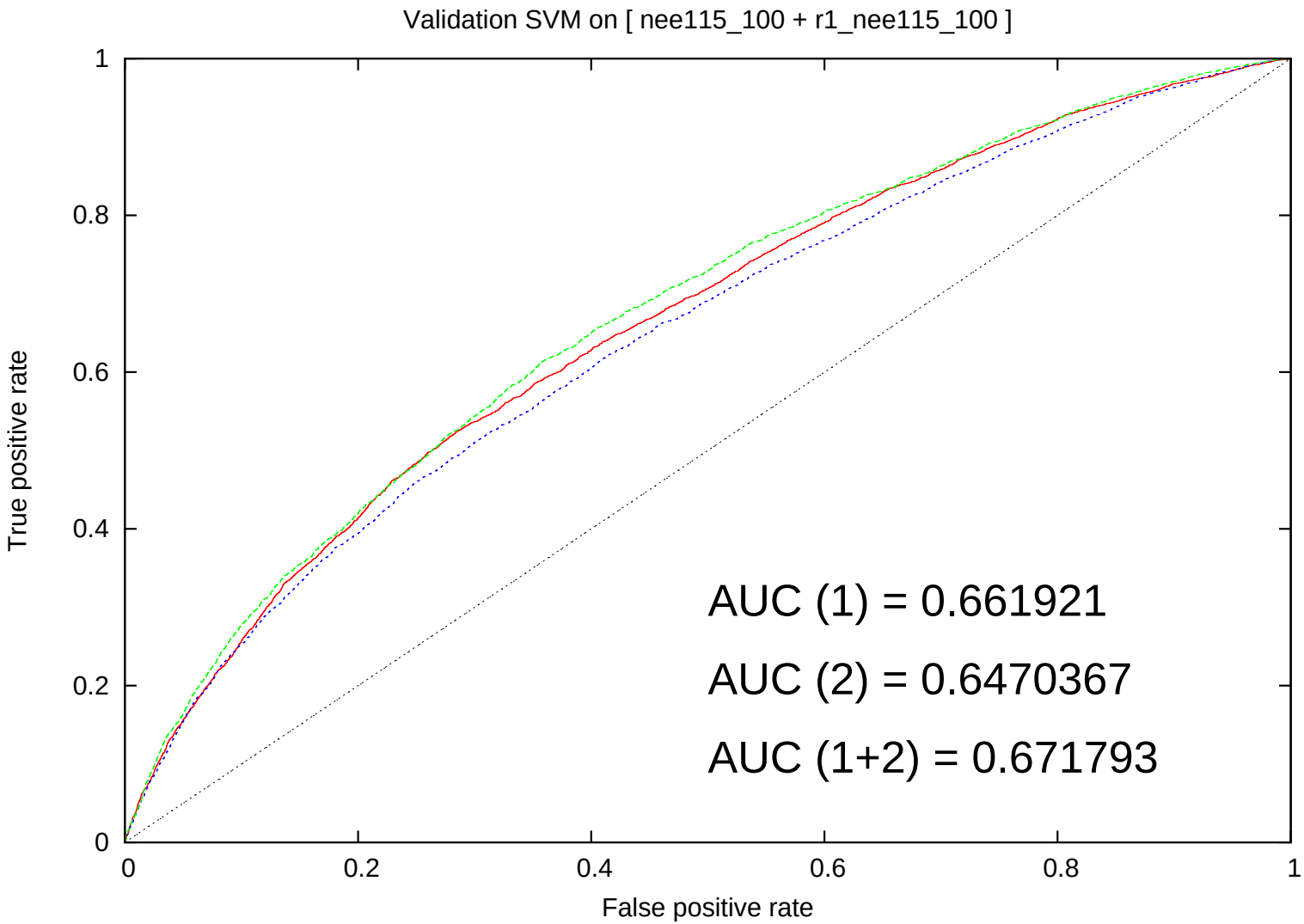




Validation of SVM on 15 different cluster numbers



Example of ROC Validation and Output combinations



Credit scoring problem and data

- 139951 clients for a building and loan credit: 3692 defaulting and 136259 non defaulting
- For each credit client: 12 variables resulting in a 23-dimensional input pattern (e.g. loan–to–value ratio, repayment rate, amount of credit, house type etc.)
- Binary target variable y: state of credit
- **Goal:** *generate (supervised or semi-supervised) binary classification functions in order to assign credit applicants to good and bad risk classes*
- In previous work **(Baesens et al. 2003, Bellotti and Crook 2008, Huang et al. 2003, Li et al. 2006, Stecking and Schebesch 2002 and 2004)** at least slightly superior performance of SVM for credit scoring towards more traditional methods could be observed

Encoding scheme for classical variables

Variable	*Scale*	*No. of Categories*	*Coding*	*Model input*
#1	nominal	3	binary 0/1	x_1 to x_2
#2	ordinal	2	binary 0/1	x_3
#3	ordinal	4	binary 0/1	x_4 to x_6
#4	nominal	2	binary 0/1	x_7
#5	nominal	5	binary 0/1	x_8 to x_{11}
#6	ordinal	5	binary 0/1	x_{12} to x_{15}
#7	nominal	3	binary 0/1	x_{16} to x_{17}
#8	nominal	3	binary 0/1	x_{18} to x_{19}
#9 – #12	quantitative	-	$\frac{x_i - \bar{x}}{s_x}$	x_{20} to x_{23}
Target	nominal	2	binary -1/1	y

Clustering large data sets

What we have before ...

- Large data set ($N \approx 140.000$), hardly feasible for advanced methods like SVM
- Relatively small set of variables
- Unequal class sizes (default rate $2,6\%$)
- Asymmetric misclassification costs
- Mixed variable types (nominal / ordinal / quantitative)
- Quantitative variables need standardization / outlier treatment

Clustering large data sets

... and after clustering

- Small data set (N = number of clusters)
- Equal class sizes (or any other desired)
- Frequency distribution, histogram, intervals, sample statistics (mean, median, standard deviation etc.) for each variable per credit client cluster
- Symbolic description of cluster objects possible $\longrightarrow$ Symbolic Data Analysis!

Clustering large data sets

What we still don't know

- Which cluster procedure to use
  - supervized / unsupervized
  - K–Means classwise
  - Constrained clustering?
- Number of clusters?
- Cluster representation?
- Cluster label (Is it a good or a bad credit client cluster?)
- How to pass cluster information to a classification method?
- How to predict credit worthiness of single credit applicants?

Cluster representation

- Let credit client i belong to cluster C_u where number of clusters is m and $C_u \in \{C_1, \ldots, C_m\}$
- Variable X_j for credit client i will take (categorical or quantitative) values X_{ij}
  $\longrightarrow$ "classical data"
- What type of characteristics X_j can we observe for cluster C_u?
  - Single valued
  - Multi valued
  - Interval valued
  - Modal multi valued
  - Modal interval valued (histogram)

  $\longrightarrow$ "symbolic data"

Clusters are symbolic objects

- Clusters are described by modal variables
- Categories or intervals appear with a given probability
- Variable X is represented by $X_u = \{\eta_{u1}, p_{u1}; \ldots; \eta_{us}, p_{us}\}$ where
  - object (cluster) u relates to variable X with $k = 1, \ldots, s$ outcomes
  - p_{uk} is probability, i.e. $\sum_{k=1}^{s} p_{uk} = 1$ and $p_k \geq 0$
  - Outcome η_{uk} may be category or (non–overlapping) interval
- For example $X_1 = \{\text{Male, } 0.65; \quad \text{Female, } 0.35\}$ might by the symbolic description of variable "gender" for cluster 1.

Encoding scheme for symbolic variables

Variable	*Original Scale*	*No. of Outcomes*	η	*Model input*
#1	nominal	3	category	p_1 to p_3
#2	ordinal	2	category	p_4 to p_5
#3	ordinal	4	category	p_6 to p_9
#4	nominal	2	category	p_{10} to p_{11}
#5	nominal	5	category	p_{12} to p_{16}
#6	ordinal	5	category	p_{17} to p_{21}
#7	nominal	3	category	p_{22} to p_{24}
#8	nominal	3	category	p_{25} to p_{27}
#9 – #12	quantitative	(4×) 4	intervals	p_{28} to x_{43}
Target	nominal	2	binary -1/1	y

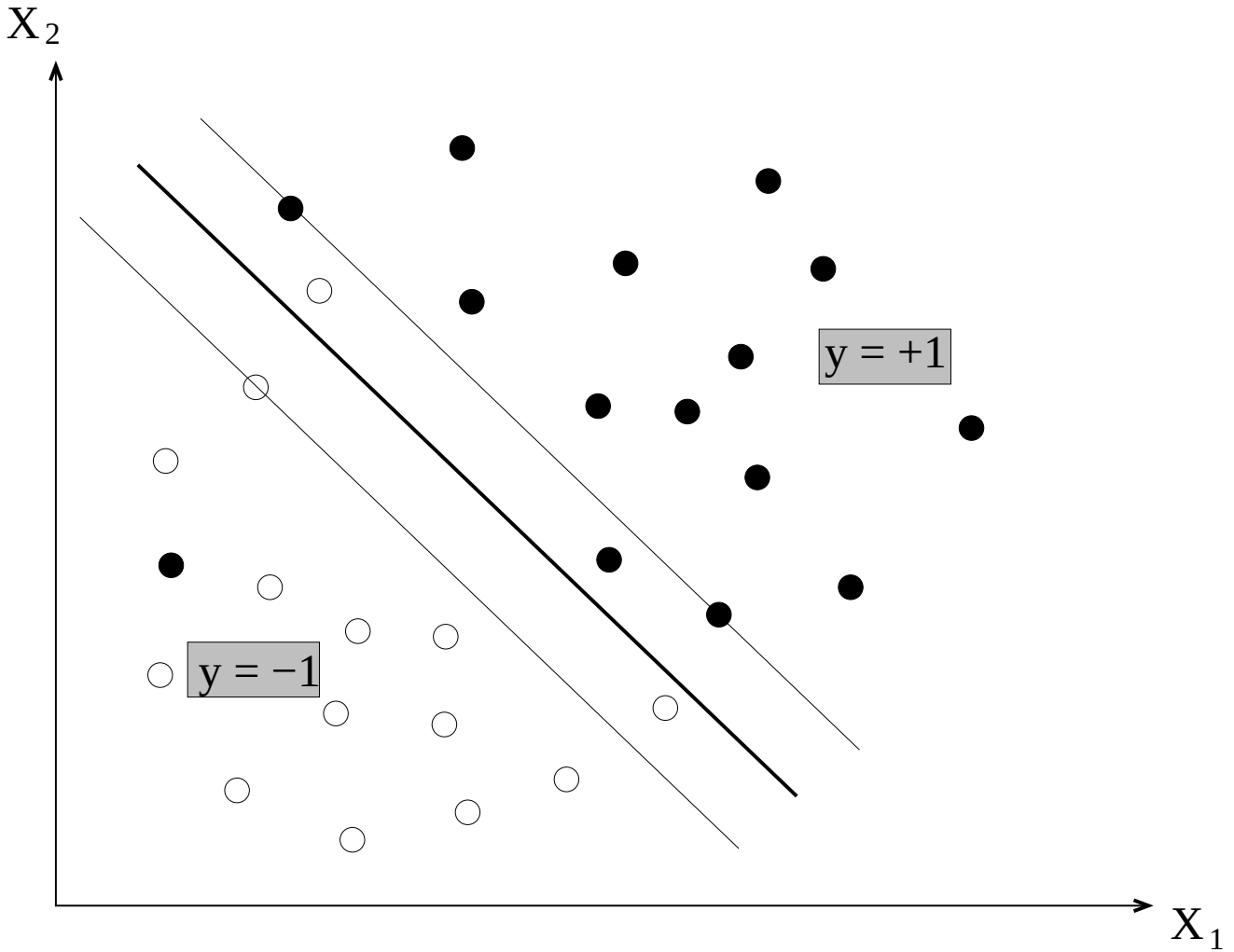
Symbolic SVM model building

Empirical approach

1. Divide large data set into "good" ($N = 136.259$) and "bad" ($N = 3.692$) credit clients
2. K–means cluster analysis: extract 50 clusters from "good" and "bad" classes respectively and preserve labels
3. Symbolic description for each cluster:
   - modal multi valued for categorical,
   - modal interval valued (with *quartiles* as interval borders) for quantitative variables
   - $\longrightarrow$ 43–dimensional input pattern consisting of probabilities is given to SVM
4. Train SVM with (a) linear and (b) RBF kernel
5. Cluster description: display region information $\longrightarrow$ Extract rules!

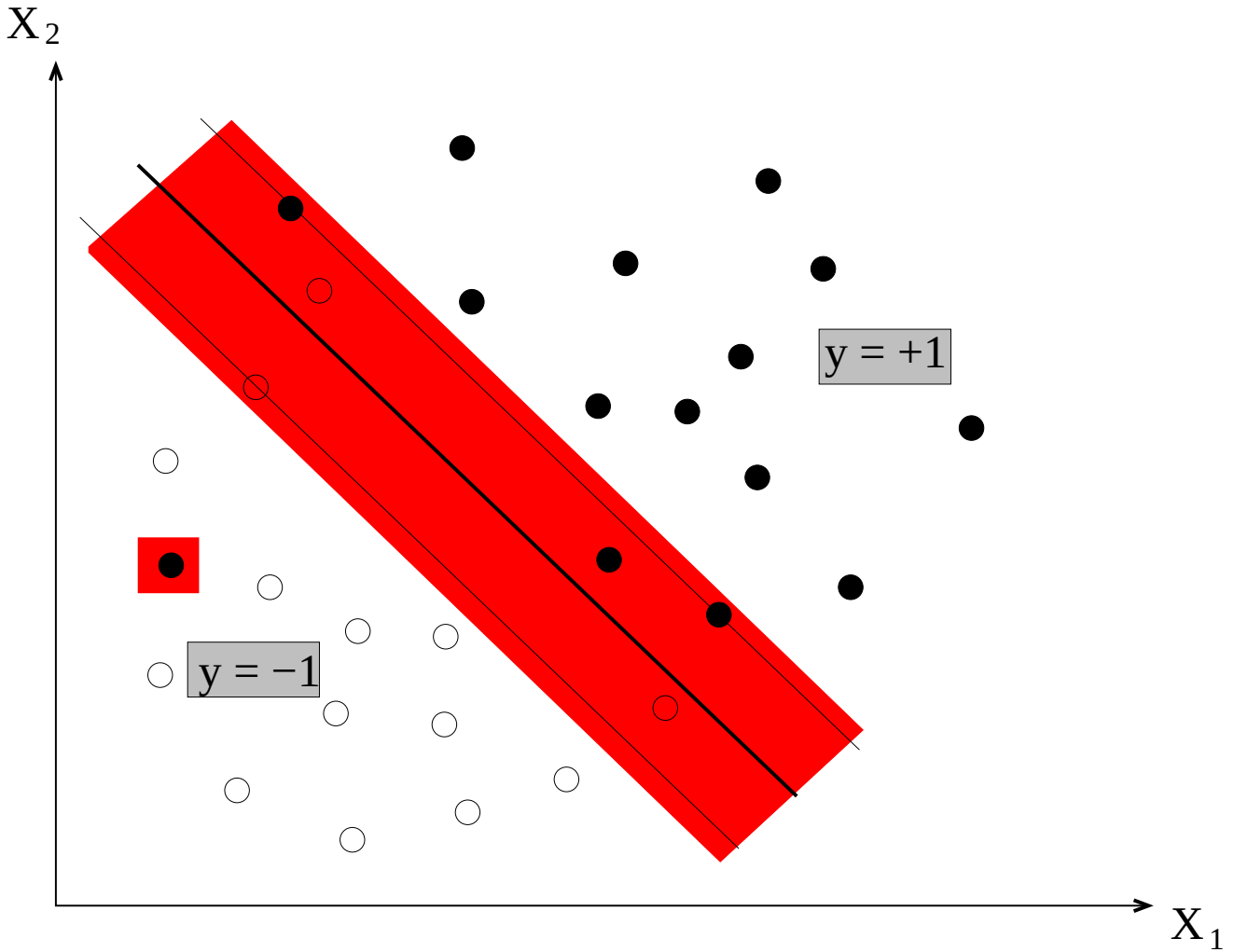
SVM classification function

Good and bad risk classes



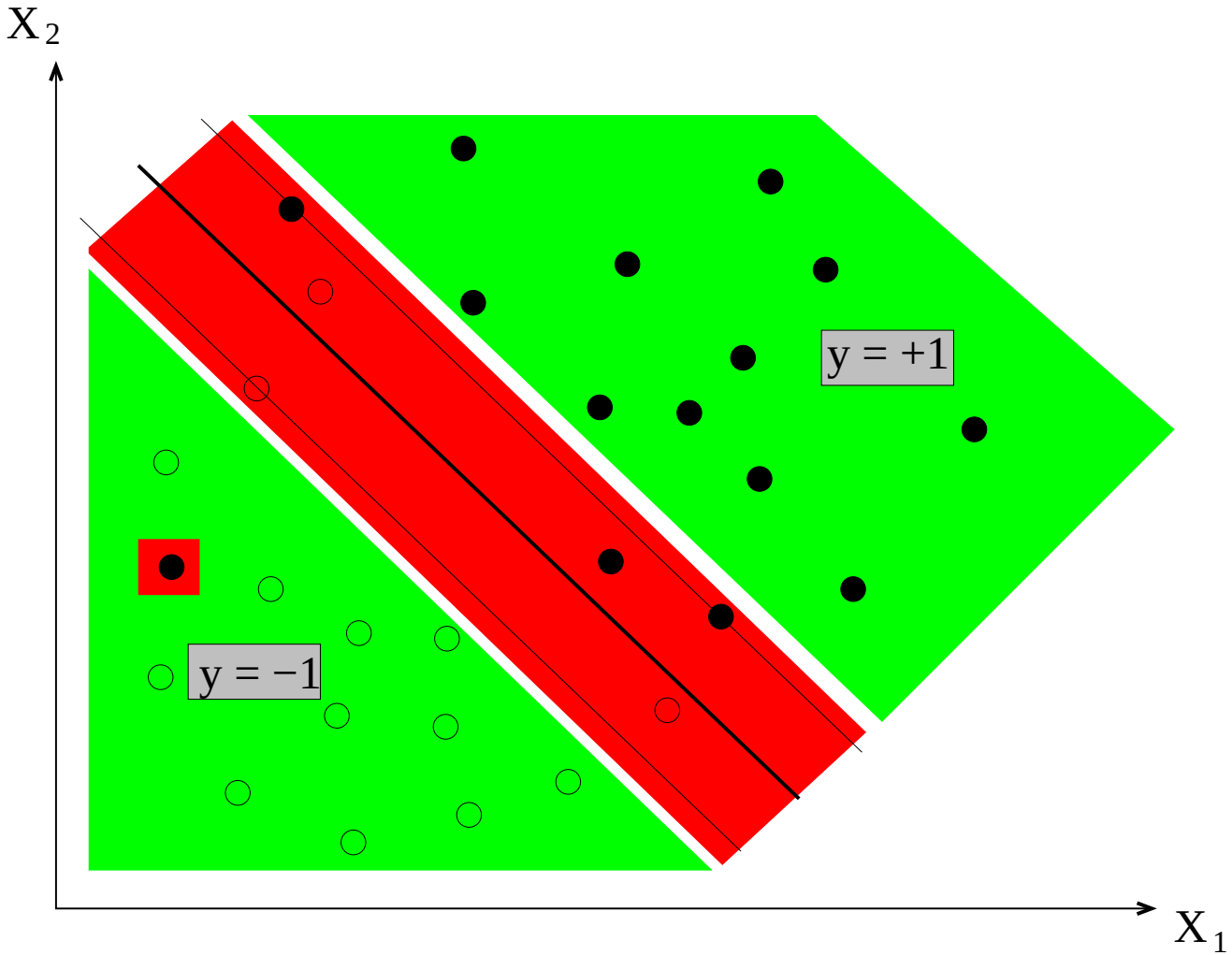
SVM classification function

Critical region



SVM classification function

Typical and critical regions



Classification with symbolic SVM

Empirical approach

1. Divide credit client data set into *training* ($N = 93.391$) and *validation* ($N = 46.560$) set
2. Cluster training set $\longrightarrow$ build SVM
3. Relative frequency coding for validation set: each credit client is assigned to a category / an interval with a probability of either 0 or 1
4. Use SVM classification function from step 2 to predict credit client behaviour on the validation set
5. Calculate ROC curve and area under ROC curve *(AUC)*
6. Compare to benchmark models: (*i*) traditional SVM and (*ii*) LDA and logistic regression trained on a small ($n = 658$) equal class sized hold out data sample

Classification results

	Cluster based SVM		Sample based SVM		Traditional Methods	
	Linear	RBF	Linear	RBF	LDA	LogReg
Training						
No. of training cases	100	100	658	658	658	658
No. of SVs	25	51	357	431	–	–
Training error	0%	0%	22.64%	10.94%	23.86%	22.49%
L–1–o error	16%	12%	27.20%	25.08%	27.66%	27.51%
Forecasting						
AUC (Full Set)						
$(N = 139951)$	0.656	0.675	0.583	0.576	0.582	0.584
[95% *CI*]	±0.09	±0.09	±0.11	±0.11	±0.10	±0.10
AUC (Validation)						
$(N = 46560)$	0.660	0.686	–	–	–	–
[95% *CI*]	±0.15	±0.15	–	–	–	–

Conclusions and outlook

- For credit scoring data with vastly different numbers of clients and with different client features, SVM trained on few cluster centers of client clusters can replicate and surpass expected (cross-validated) performance in terms of AUC.
- Clusters resulting from k-means are rather unstable on our data. This is due to overlap of or to non-spheroidal clusterings. However validated performance peaks (in terms of numbers of cluster) can be found (around 100 clusters per class).
- Cluster based SVM can be used for describing and predicting credit clients.
- Symbolic coding enables more complex representation of cluster information.
- **Further work:** on validation procedures / on more adapted symbolic cluster encodings / on data fusion