Credit Client Classification: Models Using Information from Class Boundaries and from Cluster Representatives

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

Ralf Stecking
University of Oldenburg
Department of Economics

University of Edinburgh
Credit Scoring and Credit Control XII Conference

25.08.2011

Overview

- Motivation
- Clustering and classification of credit clients
- Roles of cluster centers in classification
- Credit scoring data description and clustering
- Cluster based SVM model building and evaluation
- Conclusions and outlook

Motivation

- Past joined work during the last decade used different sized empirical credit scoring data ranging from $N = 658$ credit clients with $m = 40$ input features to $N \approx 140.000$ credit clients having $m = 25$ input features in order to analyze various classification models for credit client behavior.
- The potentially successful nonlinear SVM tend to reveal that it is difficult to identify a characteristic departure from linearity that would permit increased expected out-of-sample model performance.
- Upon observing that using cluster representatives as highly compressed training sets may not harm out-of-sample performance very much, we start to more systematically investigate to role of using such and other derived data as alternative training sets.

Clustering and classification of credit clients

- 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 $k_{ij}(x_i, x_j) \geq 0$ describes a metric relation (based on a distance) between any two training feature vectors (clients) $x_i, x_j \in \{1, ..., N\}$.
- Distances and kernels defined on client pairs are used to group (cluster) and to separate (classify) clients.
- Individualized parameters for clients pairs ij may impose conditions which act
  - class-wise (e.g. correct for asymmetric costs)
  - case-wise (e.g. correct for case importance) and
  - interaction-wise (i.e. impose 2-interactions).

Some issues concerning cluster formation

- Is there any cluster structure in the data ?
- A clustering algorithm will issue "clusters" for very $1 \leq c \leq N$!

The clustering method can function under the following conditions

- completely unsupervised
- constrained to some degree; e.g. using balancing and correlations arguments, "must-be" and "cannot-be" class assignment constraints, etc.
- constrained by labeling

How to cluster our available labeled credit client data ?

Clustering large (labeled credit client) data sets

In light of expensive distance computations ...

- Which cluster procedure to use ?
  - k-means unsupervised
  - k-means class-wise
  - other constrained clustering
- Number of clusters?
  - "Minimize" number of clusters
  - Find a "smaller" number of clusters $> m$
  - "Detail" the smaller class

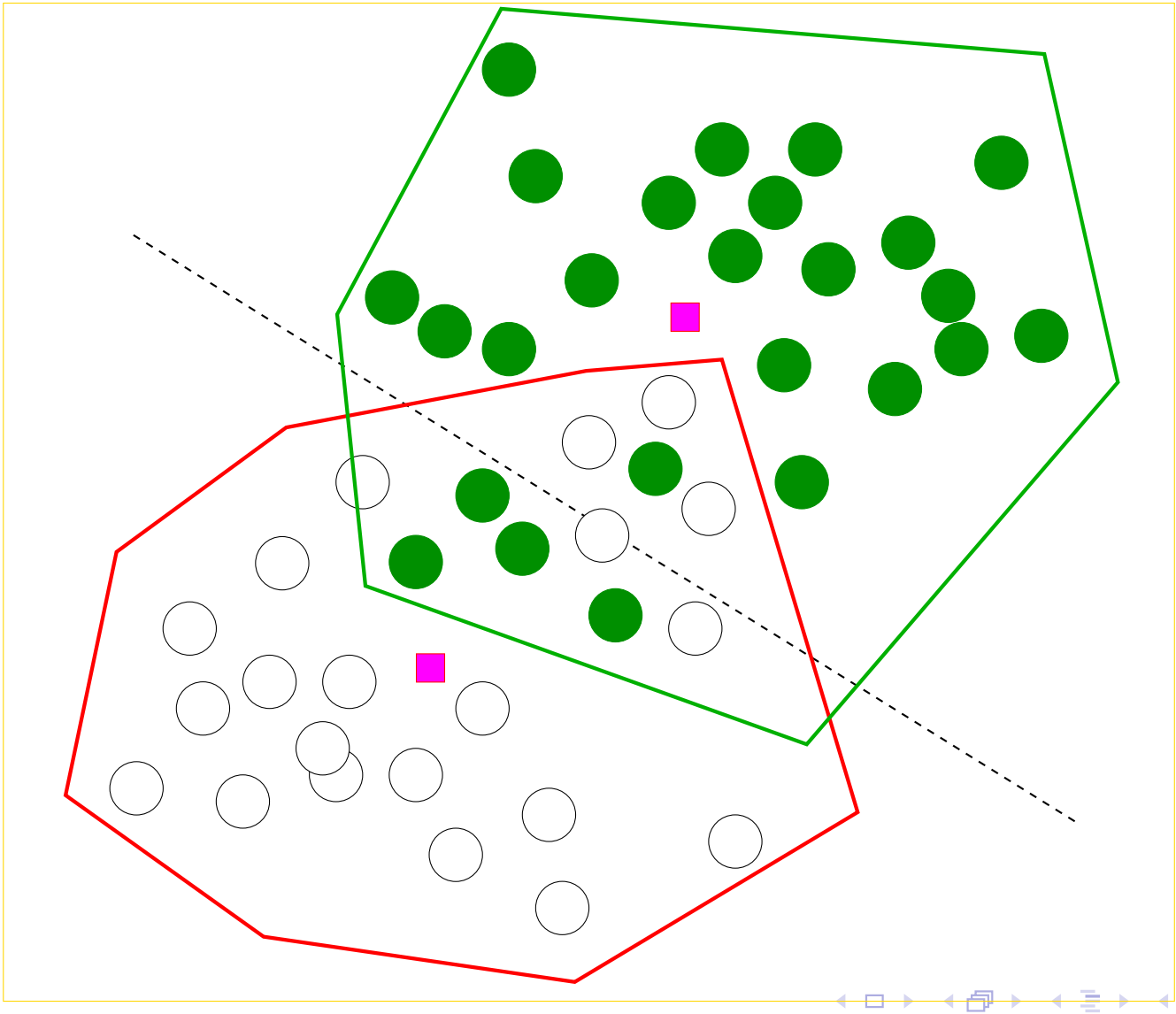
Clustering large (labeled credit client) data sets

In light of expensive distance computations ...

- Which cluster procedure to use ?
  - k-means unsupervised
  - k-means class-wise
  - other constrained clustering
- Number of clusters?
  - "Minimize" number of clusters
  - Find a "smaller" number of clusters $> m$
  - "Detail" the smaller class

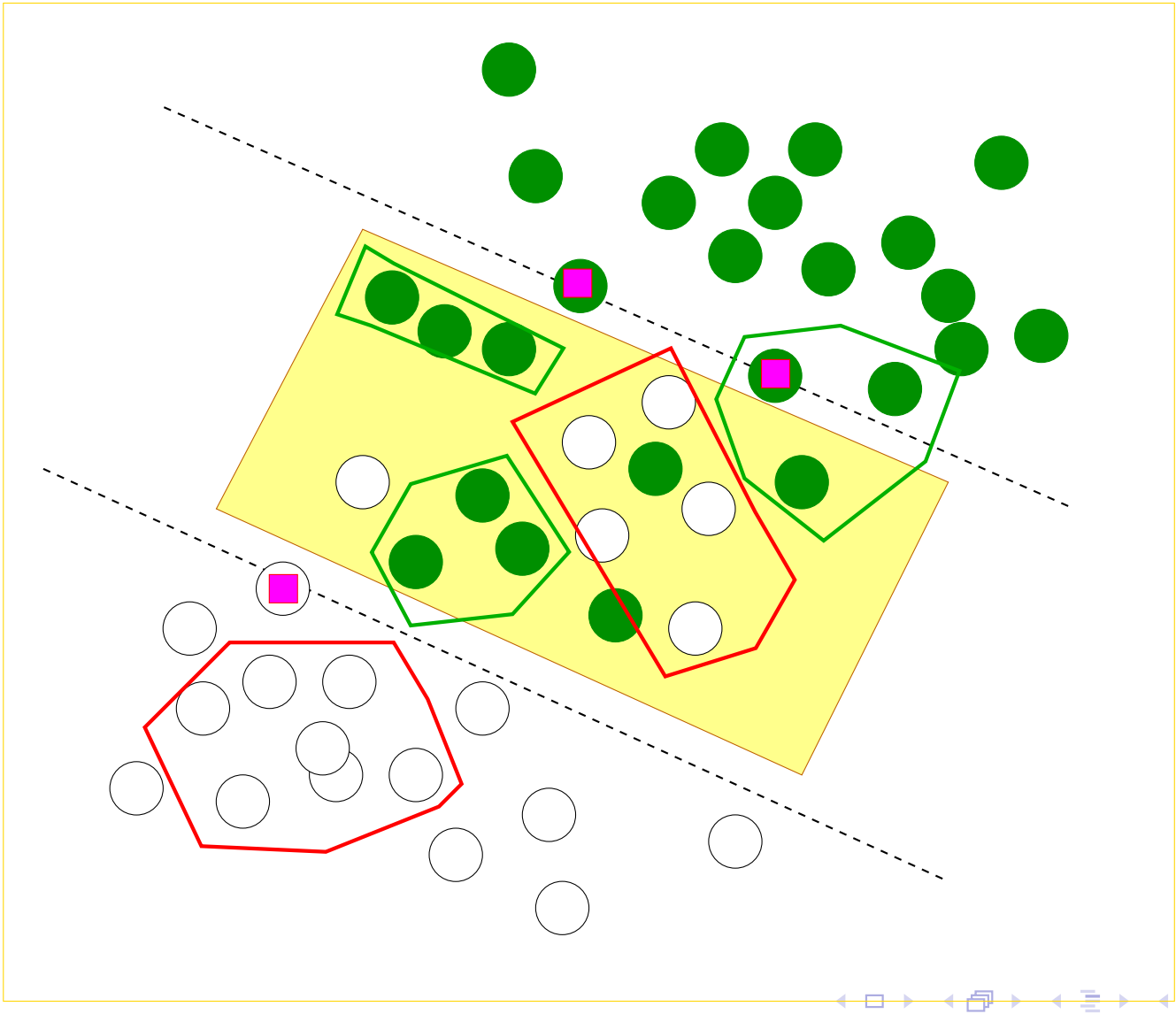
Cluster centers are class centers

... most simple use of distances ...



Cluster centers and support vectors

... i.e. when using k-means and SVM ...



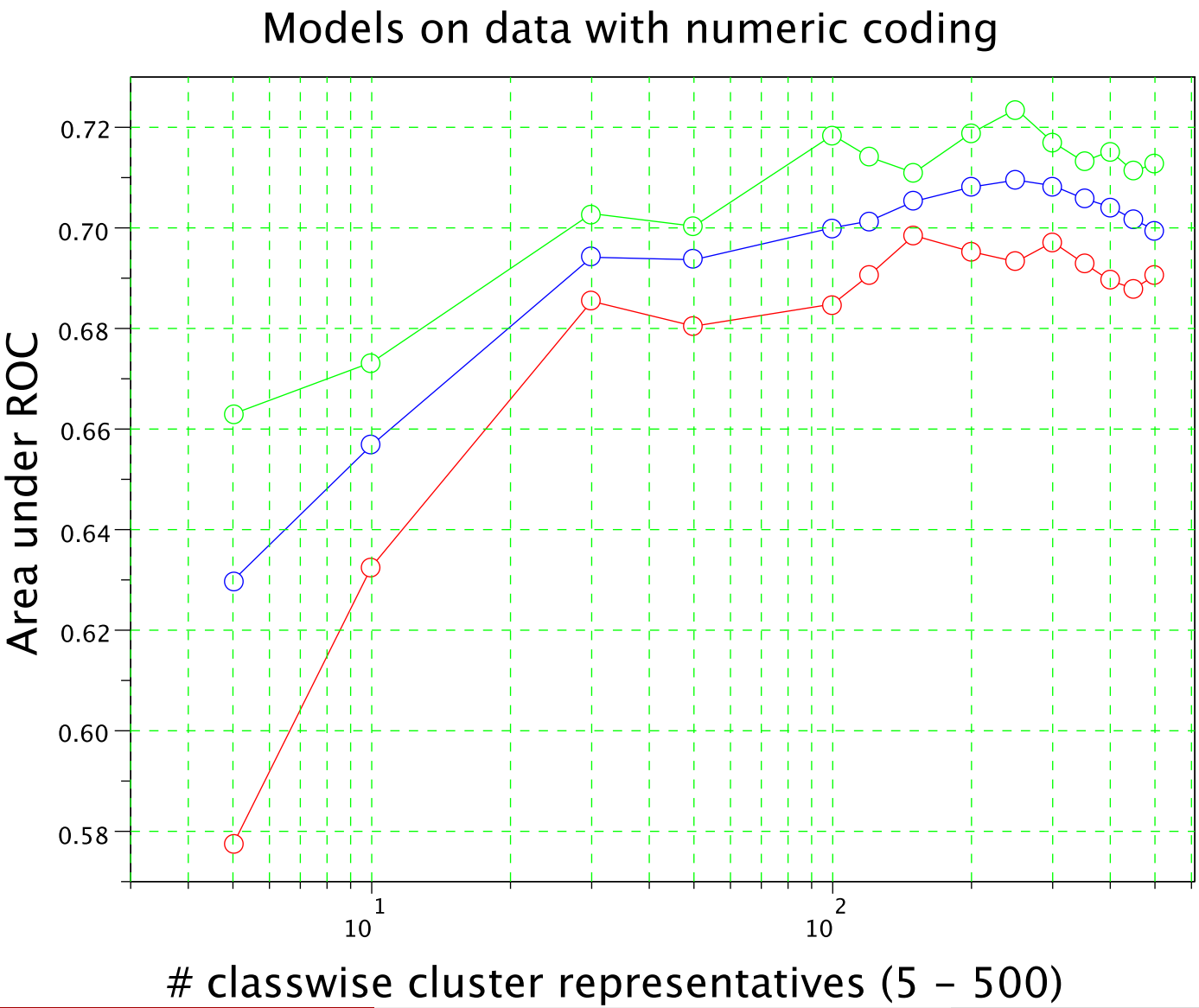
Validation of cluster based SVM on our large credit client data set

1. From the data D take positive and negative cases $D = P \cup N$.
2. Permute both P and N case wise.
3. Subdivide both P and N into n (approx. equally sized) non-overlapping segments. Set $n = 3$.
4. Cluster each set $P_1 \cup P_2$ and $N_1 \cup N_2$ thus obtaining $2k$ cluster representatives.
5. Train a SVM on these $2k$ labeled points.
6. Validate the model on the segment $P_3 \cup N_3$.

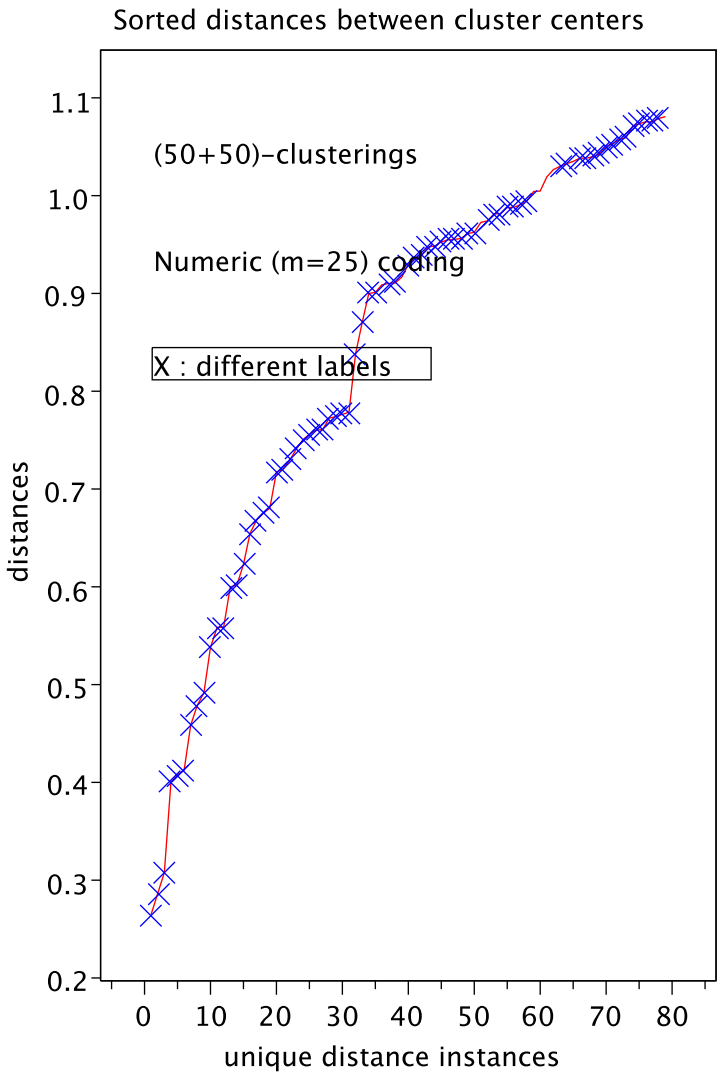
Validation of cluster based SVM on our large credit client data set ... and repeat steps ...

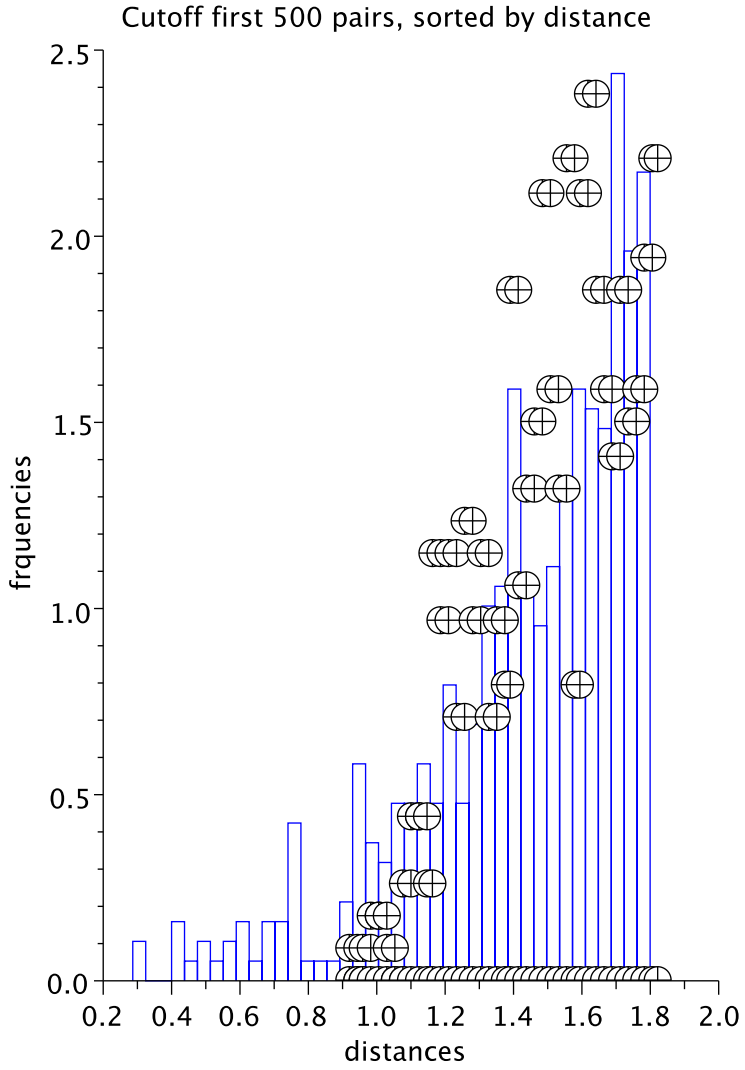
1. From the data D take positive and negative cases $D = P \cup N$.
2. Permute both P and N case wise.
3. Subdivide both P and N into n (approx. equally sized) non-overlapping segments. Set $n = 3$.
4. Cluster each set $P_1 \cup P_2$ and $N_1 \cup N_2$ thus obtaining $2k$ cluster representatives.
5. Train a SVM on these $2k$ labeled points.
6. Validate the model on the segment $P_3 \cup N_3$.

Using $m = 25$ numeric features: auROC over clusterings

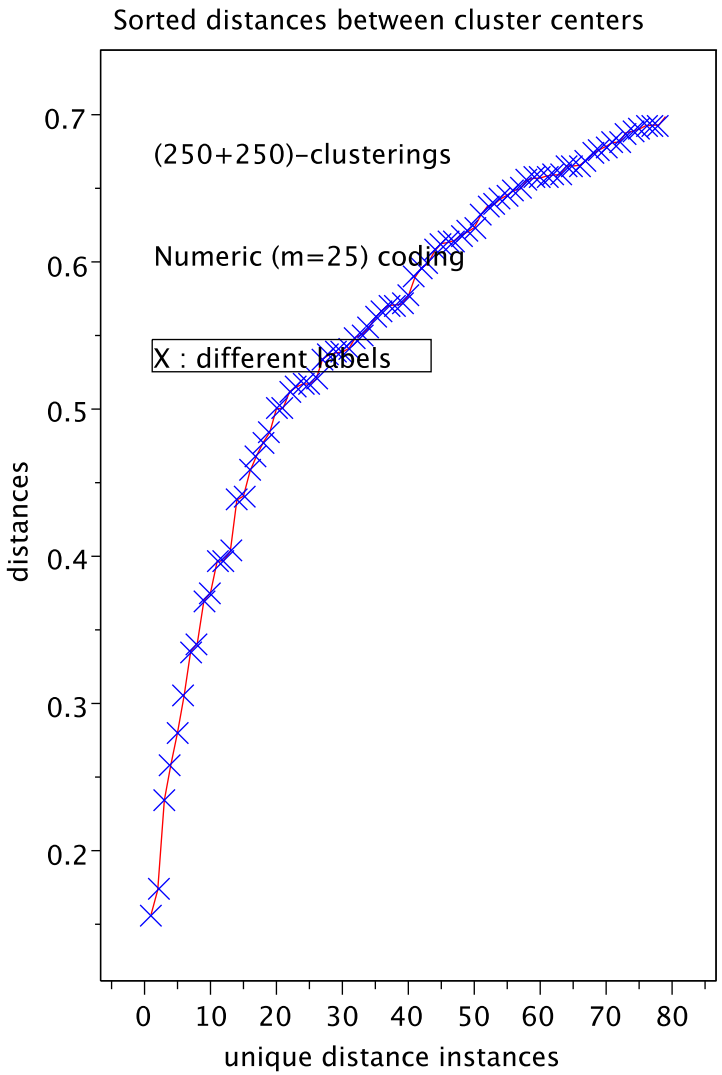


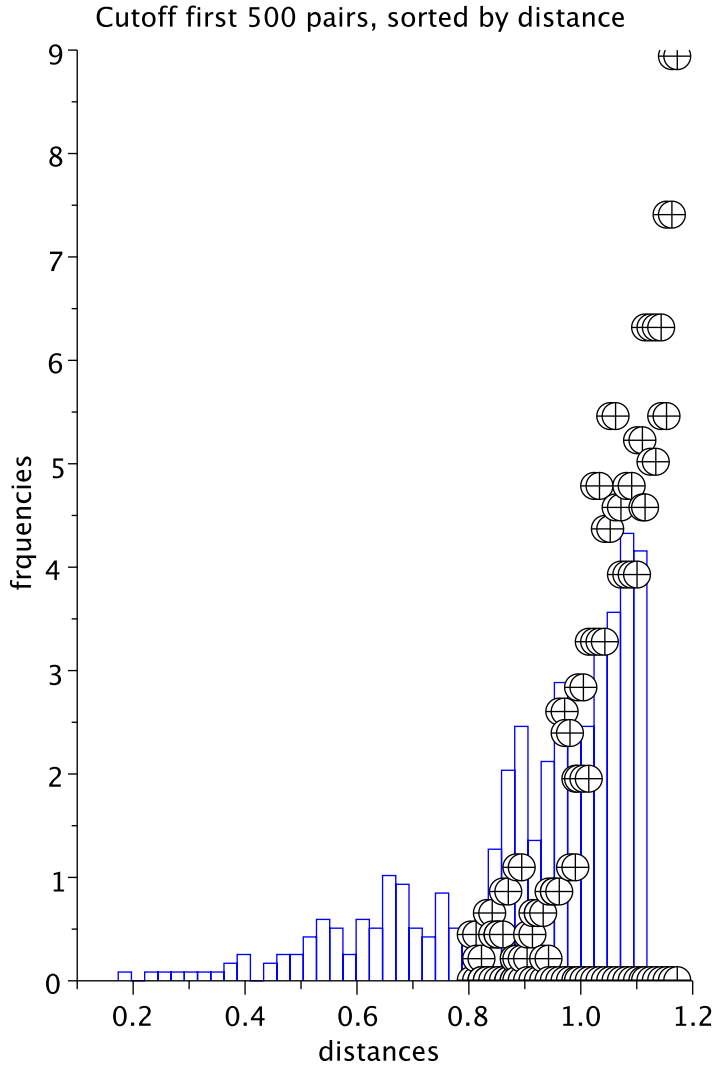
50-50 clustering: distance between cluster centers



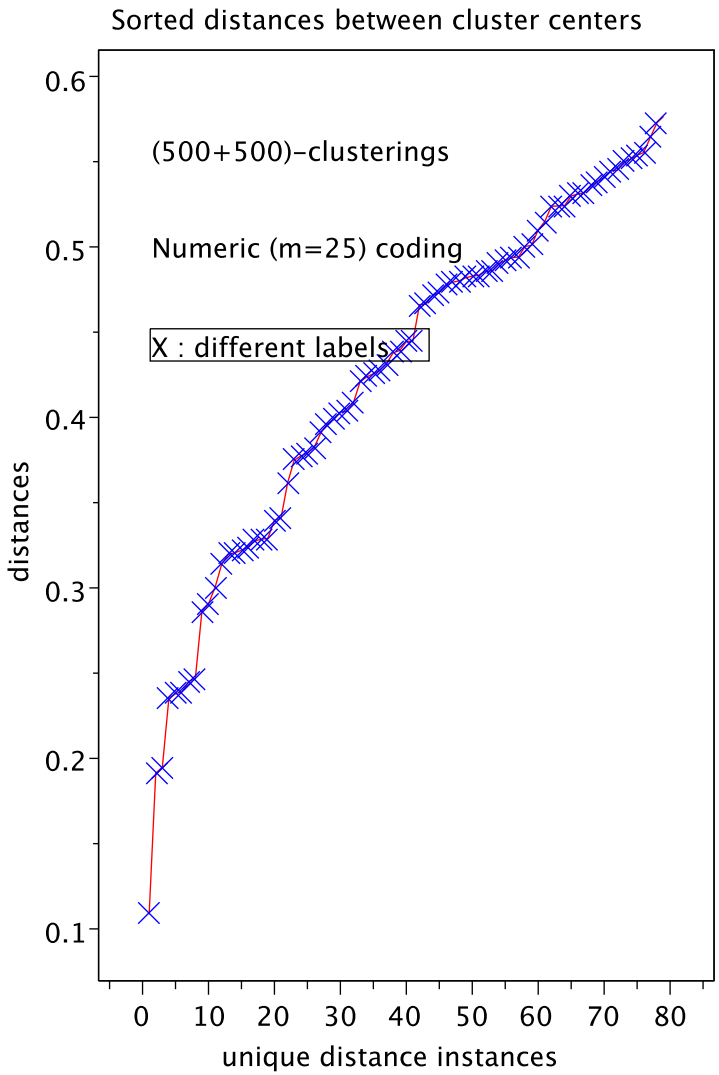


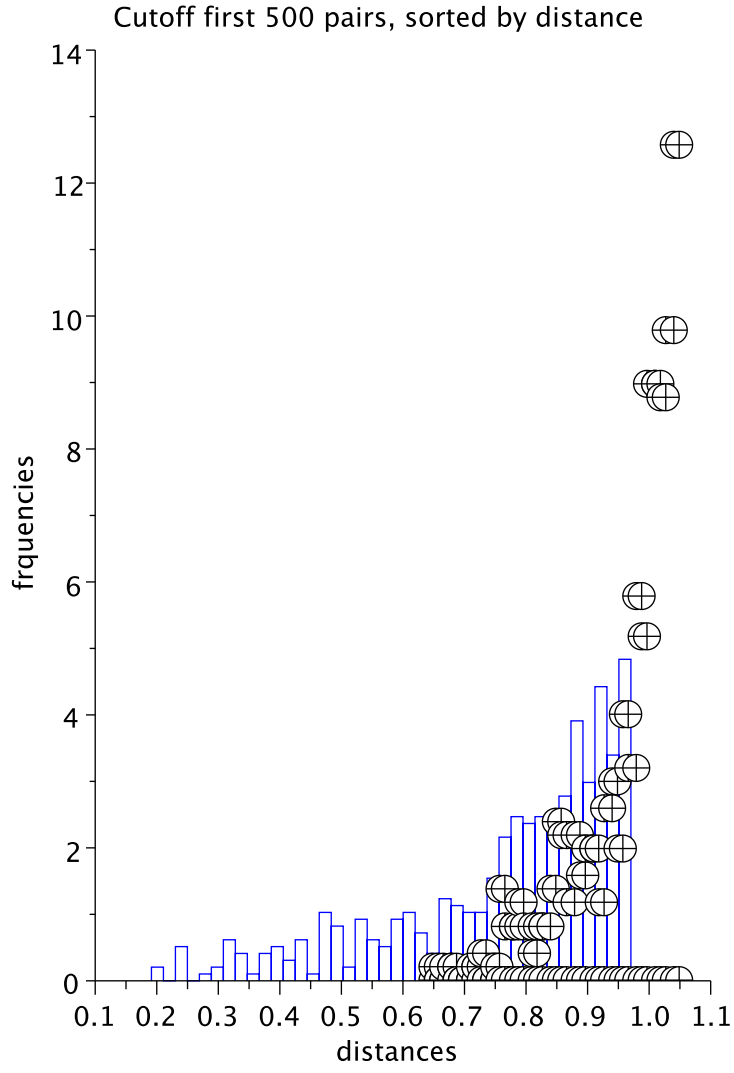
250-250 clustering: distance between cluster centers





500-500 clustering: distance between cluster centers





Credit client classification

Description of the data set

- 139951 clients for a building and loan credit
- 3692 defaulting credit clients:
  - one year default rate is 2.6%
  - binary target variable y: state of credit
- Credit client information: 12 input variables (e.g. loan–to–value ratio, repayment rate, amount of credit, house type etc.)
- Input variables are represented by a 23-dimensional input pattern

Encoding scheme for input variables

Variable	*Scale*	*No. of Categories*	*Coding*	*Model input*
#1	nominal	3	binary	x_1 to x_2
#2	ordinal	2	binary	x_3
#3	ordinal	4	binary	x_4 to x_6
#4	nominal	2	binary	x_7
#5	nominal	5	binary	x_8 to x_{11}
#6	ordinal	5	binary	x_{12} to x_{15}
#7	nominal	3	binary	x_{16} to x_{17}
#8	nominal	3	binary	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

Using cluster information for SVM

Former approaches

There are at least three approaches to combine cluster analysis and SVM:

- **Clustering training data**
  - In general: replace training set by a smaller set of (weighted) cluster representatives **Evgeniou and Pontil 2002, Shi et al. 2003 and Yu et al. 2003**
  - Large imbalanced data sets: replace examples from bigger class with cluster representatives **Yuan et al. 2006**
- **Adaptive clustering**
  - Replace non support vectors with its cluster representatives **Boley and Cao 2004**
- **Support cluster machine**
  - General extension of SVM with RBF kernel, incorporating cluster size and cluster covariance information **Li et al. 2007**

Using cluster information for SVM

Former results

- **Cluster SVM** can be used for very large data sets with millions of training examples
- Classification results for all techniques *"comparable"* or *slightly worse* than for training on the full data set **(if could be computed!)**
- Cluster techniques lead to reduced training time
- **But:** None of these techniques were used for credit scoring data

Cluster SVM model building

Experimental Setup

1. Divide credit client data set randomly into *training* ($N_T = 93.301$) and *validation* ($N_V = 46.650$) set with $N_T : N_V = 2 : 1$
2. Split training data set into "good" and "bad" credit clients
3. K–means cluster analysis: extract $\frac{n}{2}$ clusters from "good" and "bad" classes respectively and preserve labels
4. Train SVM with small data set of n cluster centers
5. Use SVM classification function to predict credit client default on the validation set
6. Calculate ROC curve and area under ROC curve *(AUC)*
   $\longrightarrow$ Go to 1. $\longrightarrow$ Repeat ten times $\longrightarrow$ Tenfold out of sample validation!

Classification results

Tenfold out of sample validation for $n = 100, \ldots, 1000$ cluster representations

No. of Clusters	*SVM RBF* ($C = 4$, $\sigma = 2.6$) *AUC (Validation Set,* $N = 46650$*)*			
	Mean	*Std. Dev.*	*Minimum*	*Maximum*
100	0.699	0.007	0.687	0.708
200	0.705	0.009	0.686	0.716
300	0.708	0.006	0.696	0.717
400	0.712	0.009	0.699	0.728
500	0.712	0.008	0.701	0.722
600	0.710	0.008	0.691	0.718
700	0.708	0.009	0.688	0.721
800	0.708	0.008	0.693	0.722
900	0.707	0.009	0.696	0.726
1000	0.707	0.007	0.696	0.716

Full data and sampled data SVM model building

Experimental Setup

1. Divide credit client data set randomly into *training* ($N_T = 93.301$) and *validation* ($N_V = 46.650$) set with $N_T : N_V = 2 : 1$
2. Sample 500 clients from "good" and "bad" training examples each
   $\longrightarrow$ store sampled training data
3. Train SVM with
   (a) full data training set from 1.
   (b) sampled data training set from 2.
4. Use SVM classification functions from 3(a) and 3(b) to predict credit client default on the validation set
5. Calculate ROC curve and area under ROC curve *(AUC)*
   $\longrightarrow$ Go to 1. $\longrightarrow$ Repeat ten times $\longrightarrow$ Tenfold out of sample validation!

Classification results

Tenfold out of sample validation for full data and sampled data set

	Full Data based SVM	Sampled Data based SVM	Cluster based SVM
Training			
Kernel Type	Linear	RBF	RBF
Parameters	$C = 100$	$C = 5,\ \sigma = 4.5$	$C = 4,\ \sigma = 2.6$
No. of training examples	93301	1000	500
Validation			
No. of validation examples	46650	46650	46650
AUC			
Mean	0.702	0.702	0.712
Std.Dev.	0.005	0.008	0.008
Min.	0.695	0.693	0.701
Max.	0.707	0.719	0.722

Conclusions and outlook

- Class-wise clustering of training data by k-means leads to credit client classification models with peak expected performance at a medium number of clusters.
- This holds for different data encodings.
- Sample based *(non-linear)* SVM does **not** lead to superior classification results when compared to *(linear)* SVM trained on the full data set
- Cluster based SVM classification results are *slightly but significantly* **better** than sample based SVM (significance level $p = 0.018$) and full set SVM ($p = 0.014$)
- **Further work:** on constrained clustering / on more adapted symbolic cluster encodings / on data fusion