How to select representative data to train SVM efficiently?

Xiaoou Li

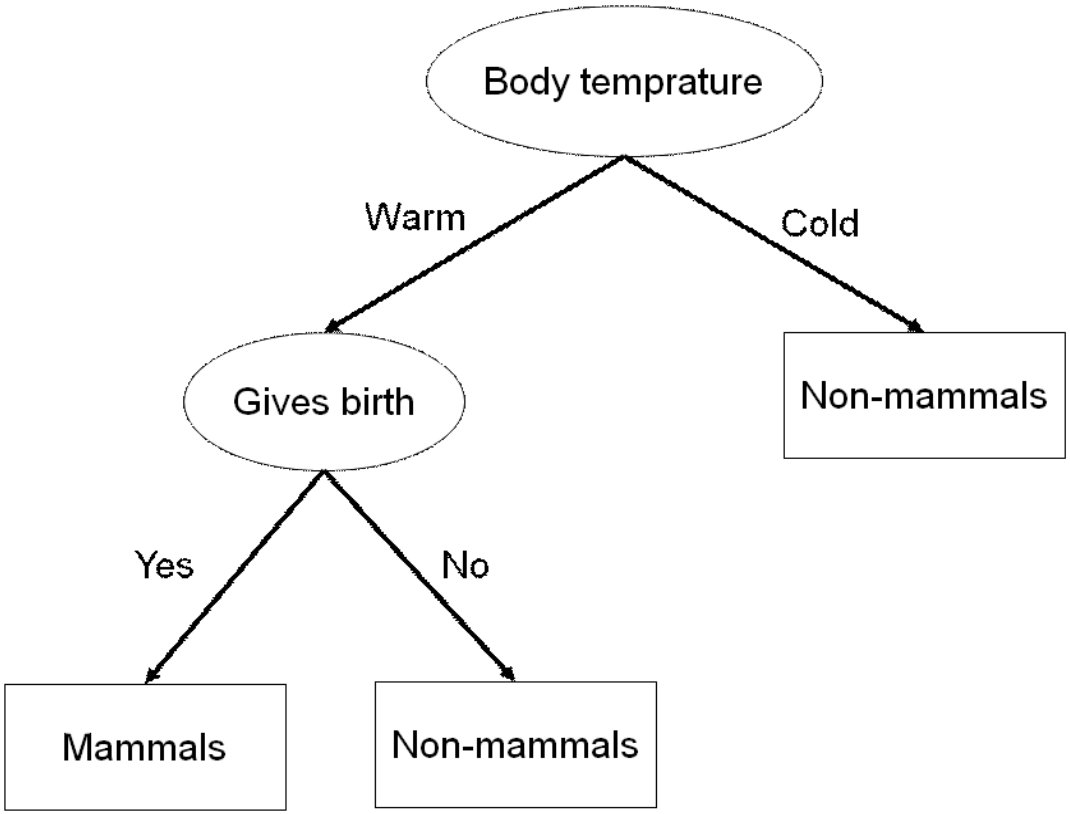
Departamento de Computación
Centro de Investigación y de Estudios Avanzados del Instituto Politécnico Nacional (CINVESTAV-IPN)
México

Outline

Classification-

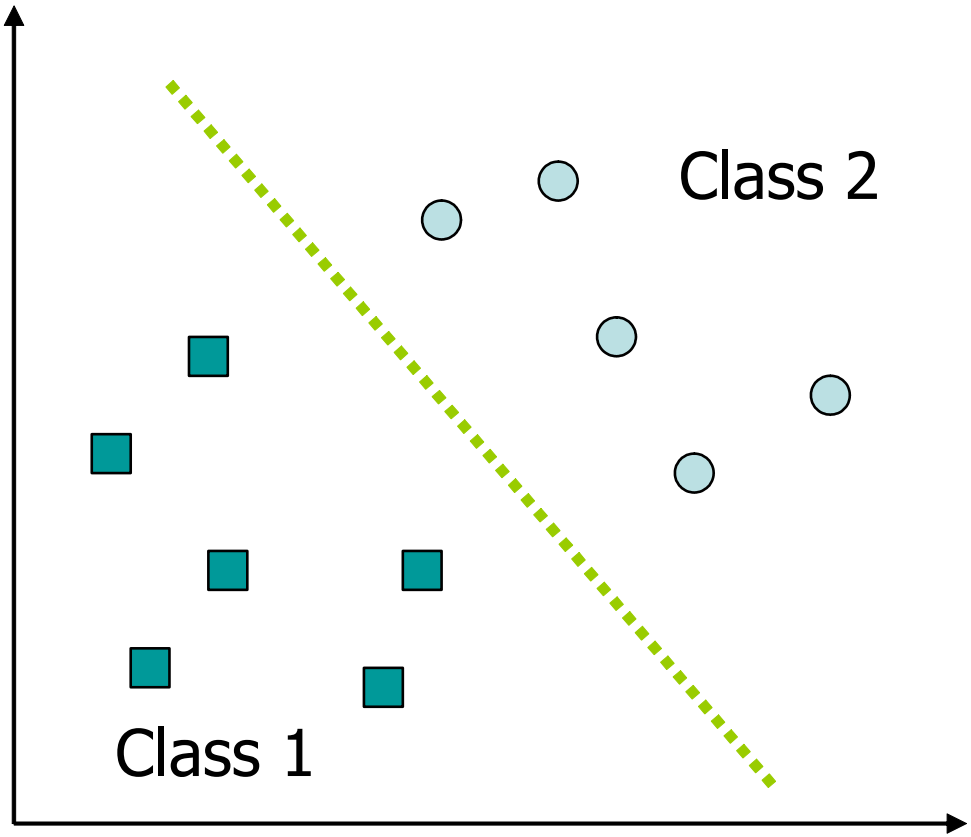
Definition

Classification is the process of **predicting** the class of given data points. Classes are sometimes called as *targets/ labels or categories*. Classification predictive modeling is the task of **approximating** a mapping function (f) from input variables (X) to discrete output variables (y).



Classification-

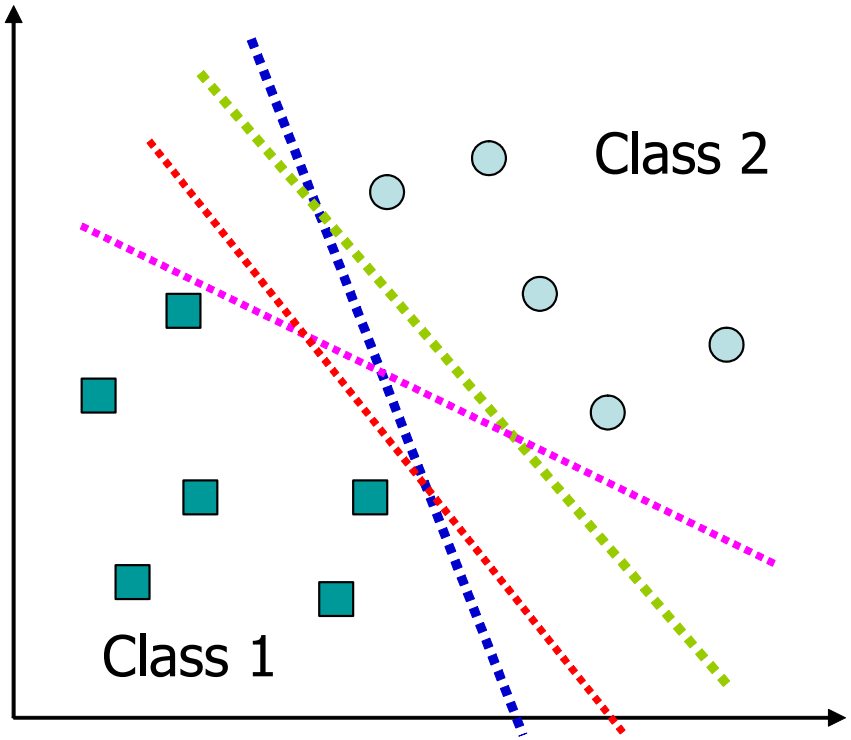
Consider a two-class, linearly separable classification problem.
Let $\{x_1, ..., x_n\}$ be our data set and let $y_i \in \{1, -1\}$ be the class label of x_i



Classification-

What is a good Decision Boundary?

Many decision boundaries!

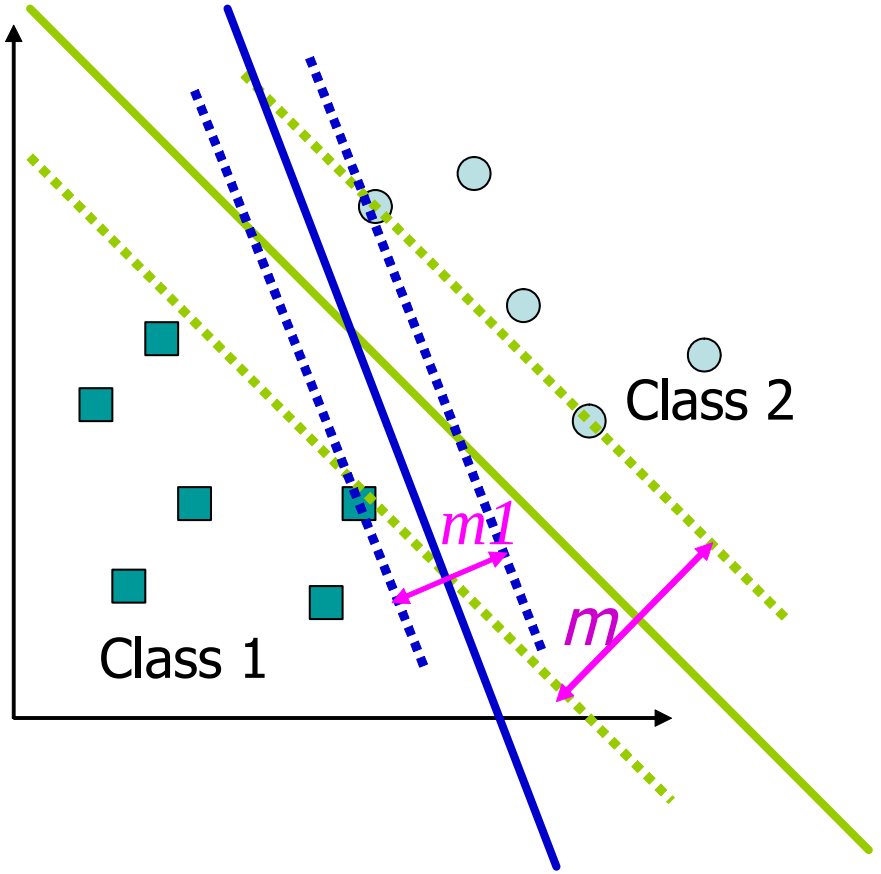


Are all decision boundaries equally good?

— Many algorithms have been proposed, Rule based, Nearest-neighbor, Naive Bayes, Artifitial neural networks(ANN), Support vector machine(SVM).....

Classification-

Examples of Bad Decision Boundaries



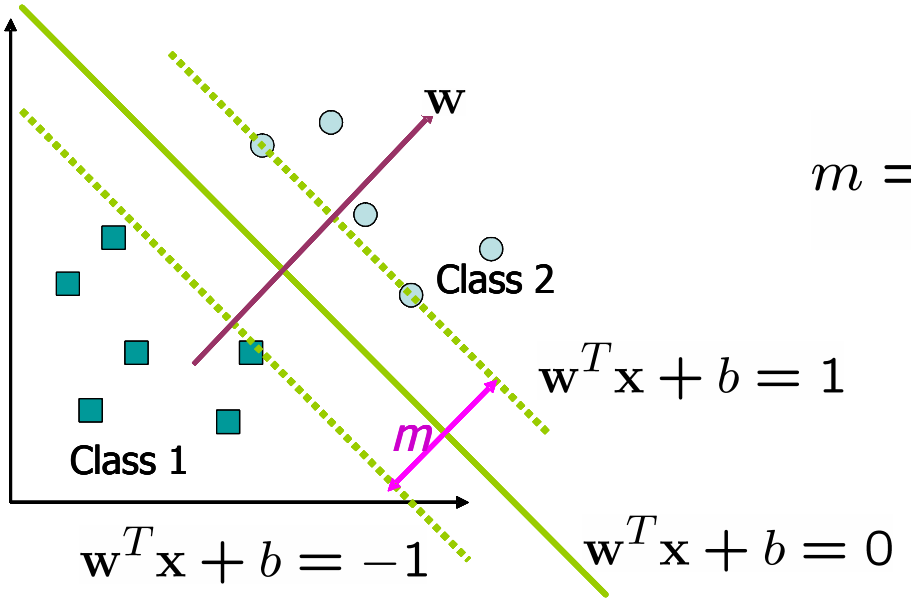
Which one is better?

Classification-linear case

Large-margin Decision Boundary

The decision boundary should be as far away from the data of both classes as possible

We should maximize the margin *m*, *maximal margin hyperplane*



$$m = \frac{2}{||\mathbf{w}||}$$

Classification-linear case

Finding the Decision Boundary

The decision boundary should classify all points correctly, i.e.,
$y_k \left[w^T x_k + b \right] \geq 1$
The decision boundary can be found by solving the following constrained optimization problem

$$\begin{aligned} &\text{Minimize } \frac{1}{2} \|w\|^2 \\ &\text{subject to : } y_k \left[w^T x_k + b \right] \geq 1 \end{aligned}$$

The decision boundary should classify all points correctly, i.e.,
$y_k \left[w^T \varphi\left(x_k\right) + b\right] \geq 1$
The decision boundary can be found by solving the following constrained optimization problem

$$\begin{aligned} &\text{Minimize } \frac{1}{2}\left\|w\right\|^2 \\ &\text{subject to : } y_k \left[w^T \varphi\left(x_k\right) + b\right] \geq 1 \end{aligned}$$

This is a constrained optimization problem. Solving it requires some new tools, $\varphi\left(x_k\right)$ is a nonlinear function.

$$\begin{aligned} &\min_{w,b} J(w) = \tfrac{1}{2} w^T w + c \sum_{k=1}^{n} \xi_k \\ &\text{subject} : y_k \left[w^T \varphi(x_k) + b \right] \geq 1 - \xi_k \end{aligned} \tag{1}$$

where ξ_k is slack variable to tolerate mis-classifications $\xi_k > 0$, $k = 1 \cdots n$, $c > 0$, $\frac{1}{\|w_k\|}$ is the distance from x_k to the hyperplane $\left[w^T \varphi(x_k) + b \right] = 0$. (1) is equivalent to the following dual problem with the Lagrangian multipliers $\alpha_k \geq 0$

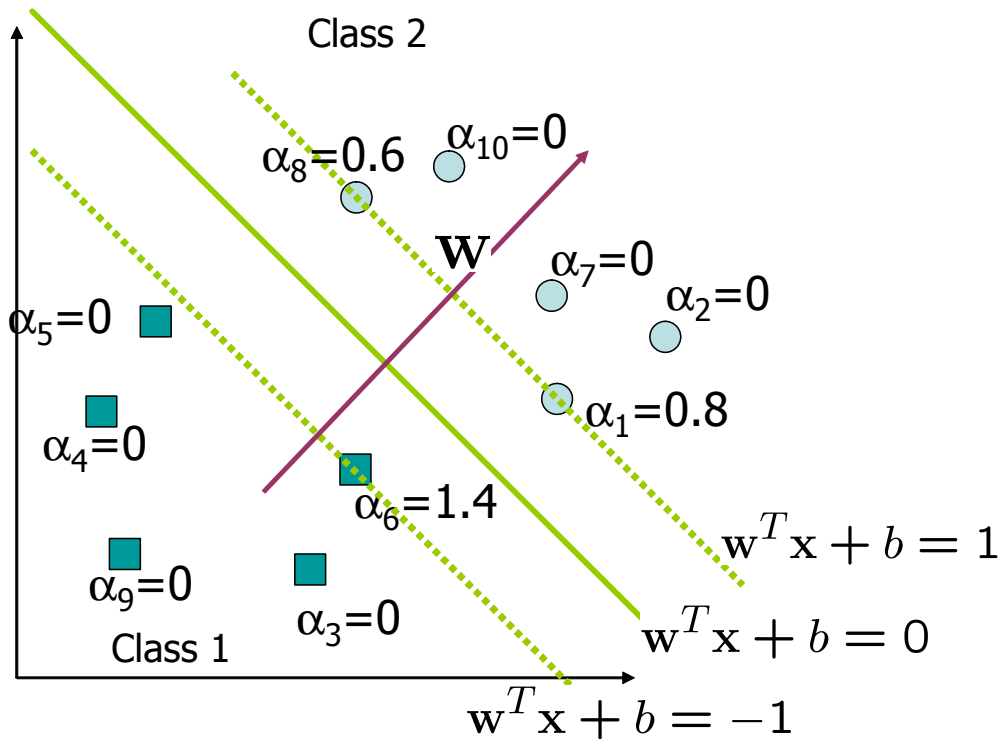
$$\begin{aligned} &\max_{\alpha} J(\alpha) = -\tfrac{1}{2} \sum_{k,j=1}^{n} y_k y_j K(x_k, x_j) \alpha_k \alpha_j + \sum_{k=1}^{n} \alpha_k \\ &\text{subject} : \sum_{k=1}^{n} \alpha_k y_k = 0, \quad 0 \leq \alpha_k \leq c \end{aligned} \tag{2}$$

where the kernel $K(x_k, x_i)$ satisfies the Mercer condition: $K(x_k, x_i) = \varphi(x_k)^T \varphi(x_i)$.

- Many of the α_i are zero
  - W is a linear combination of a small number of data points: $W = \sum_{j=1}^{s} \alpha_{t_j} y_{t_j} X_{t_j}^T$, where $s \ll n$.
- X_i with non-zero α_i are called support vectors (SV)
  - The decision boundary is determined only by the SV
- For testing with a new data z
  - Compute $W^T z + b = \sum_{j=1}^{s} \alpha_{t_j} y_{t_j} (X_{t_j}^T z) + b$ and classify z as class 1 if the sum is positive, and class 2 otherwise.

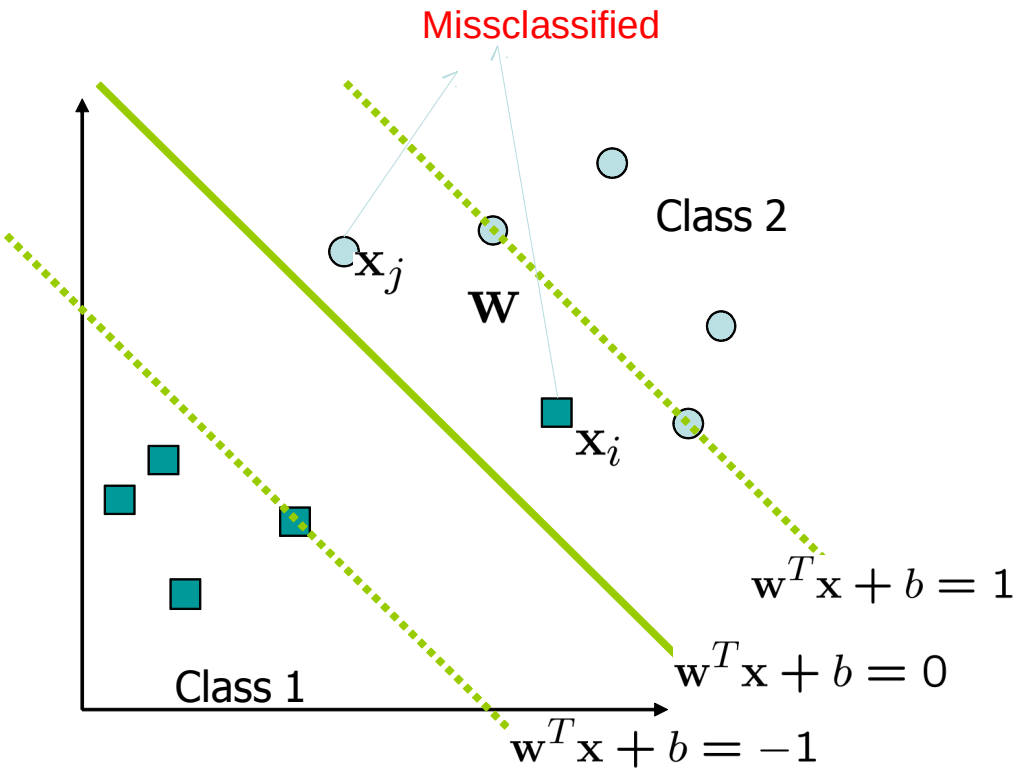
SVM-

A Geometrical Interpretation



SVM-

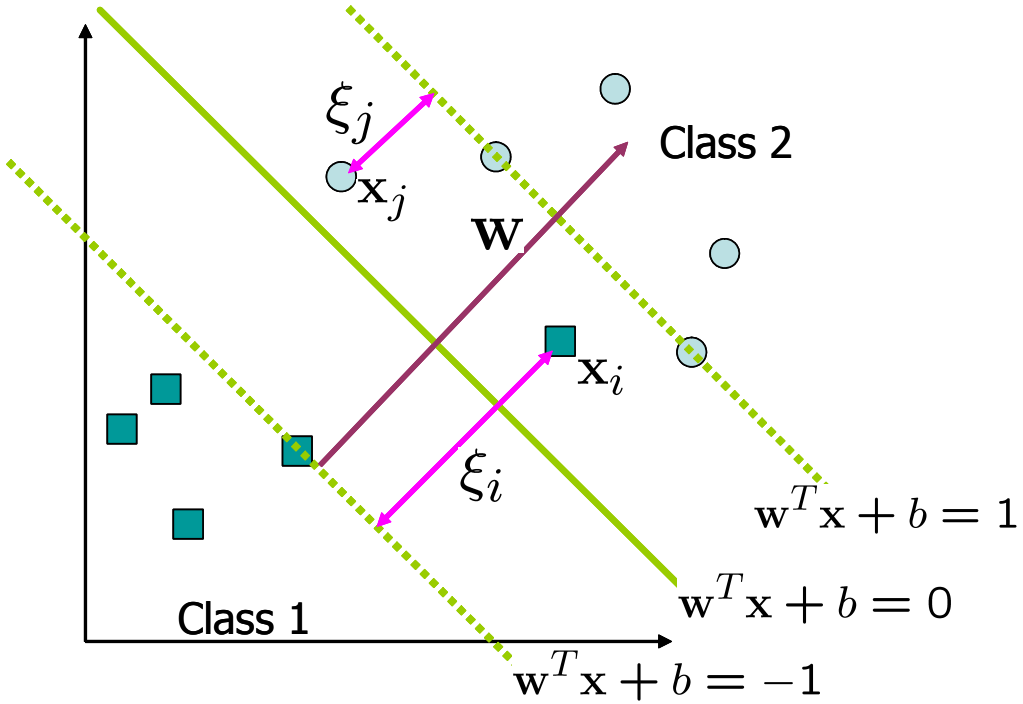
Non-linearly Separable Problems



SVM-

Non-linearly Separable Problems

- We allow "error" ξ_i in classification; it is based on the output of the discriminant function $W^T X + b$
- ξ_i approximates the number of misclassified samples

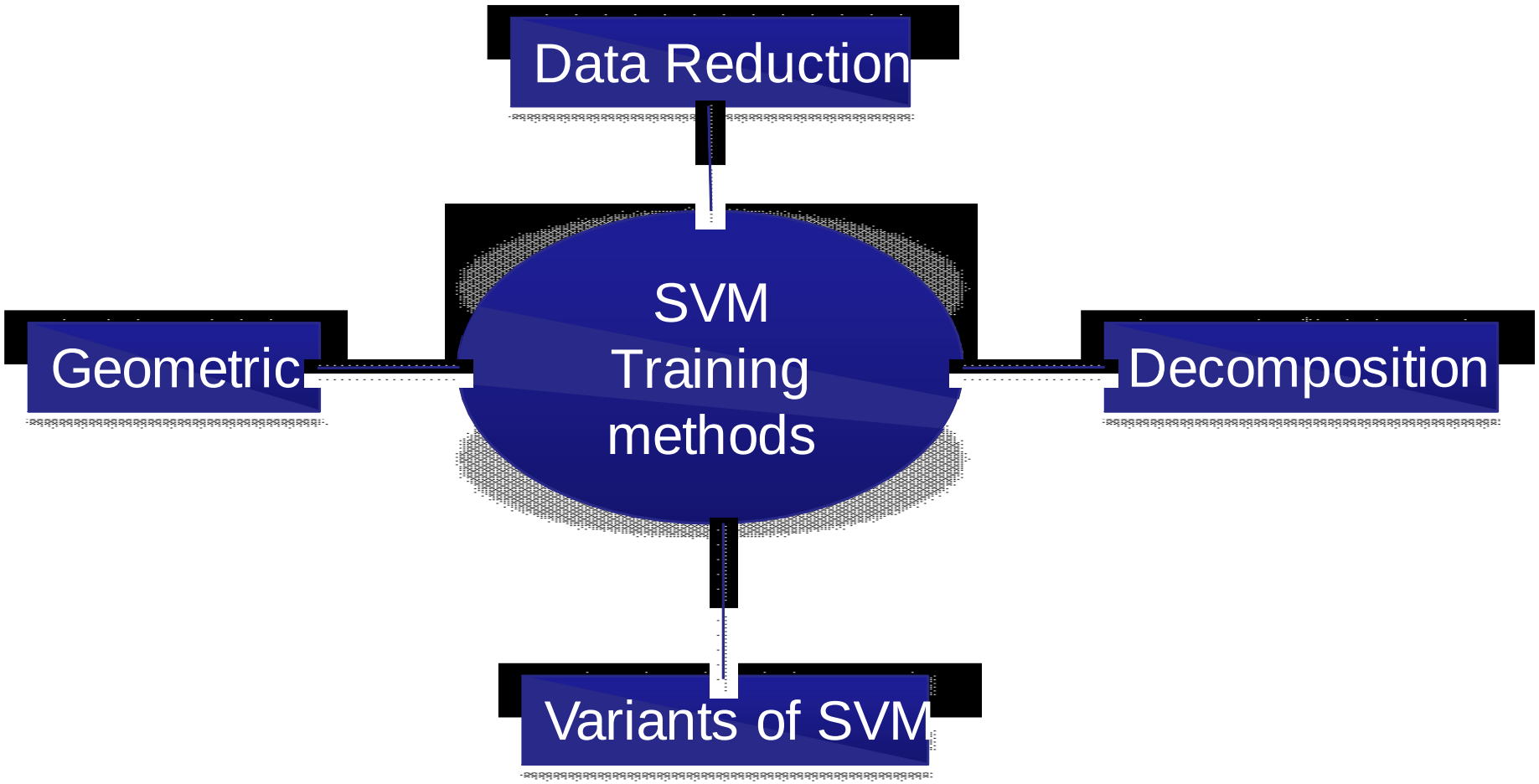


- When the the data n is large, it is difficult to solve the QP problem (2). Sequential minimal optimization (SMO) is one of the most popular method.
  - A QP with two variables is trivial to solve
  - Each iteration of SMO picks a pair of (α_i,α_j) and solve the QP with these two variables; repeat until convergence
- In practice, we can just regard the QP solver as a "black-box" without bothering how it works

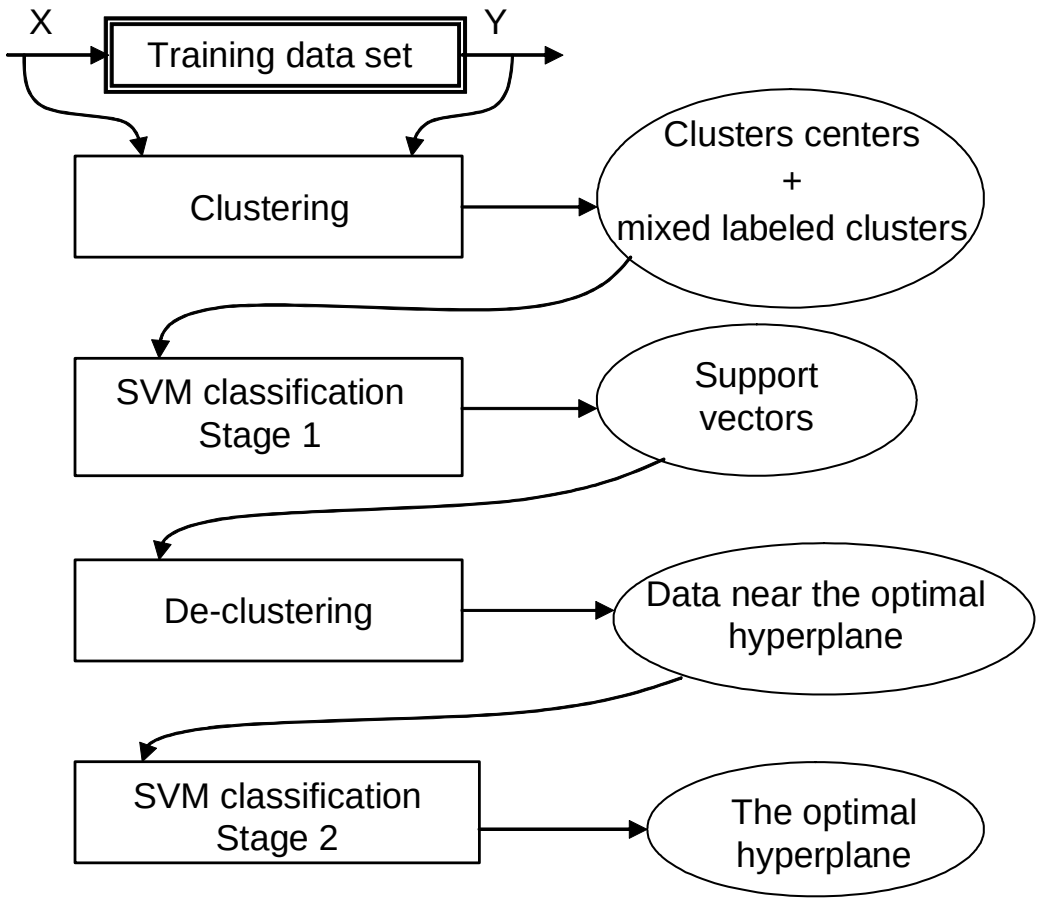
Large data set classification

- Modify SVM so that it could deal with large data sets within an acceptable time
  - projected conjugate gradient (PCG) chunking algorithm
  - Sequential Minimal Optimization (SMO)
  - parallel optimization step
  - Genetic programming
  - Neural networks
- Reduce a large data set to a smaller one so that normal SVM algorithms could be applied
  - Clustering, e.g., hierarchical clustering, k-means cluster, parallel clustering
  - Rocchio bundling
  - Bayesian committee machine
  - Random selection

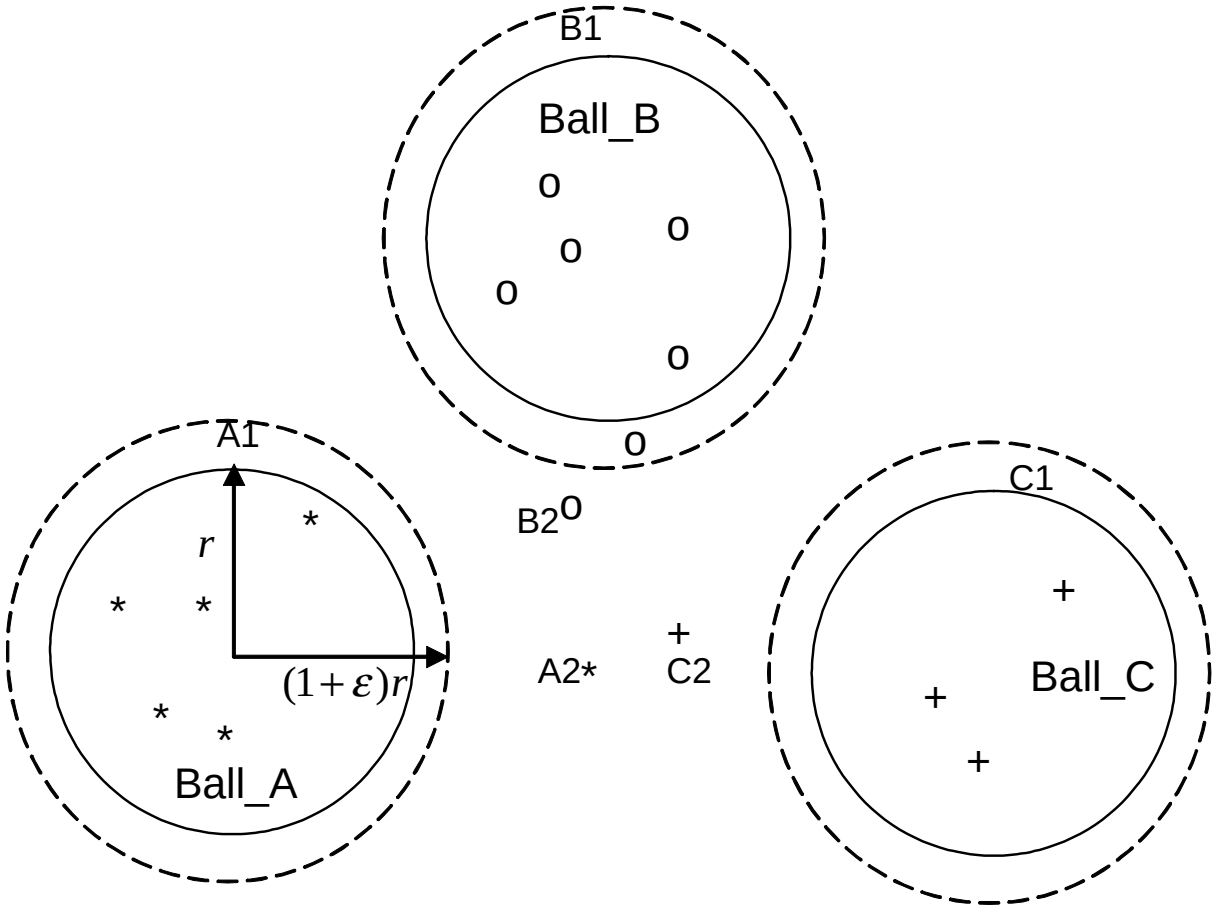
SVM training methods



)


X
Training data set
Y
Clustering
Clusters centers
+
mixed labeled clusters
SVM classification
Stage 1
Support
vectors
De-clustering
Data near the optimal
hyperplane
SVM classification
Stage 2
The optimal
hyperplane

Minimum enclosing ball (MEB) clustering



MEB-SVM^2

Two-stage classification via minimum enclosing ball (MEB) clustering

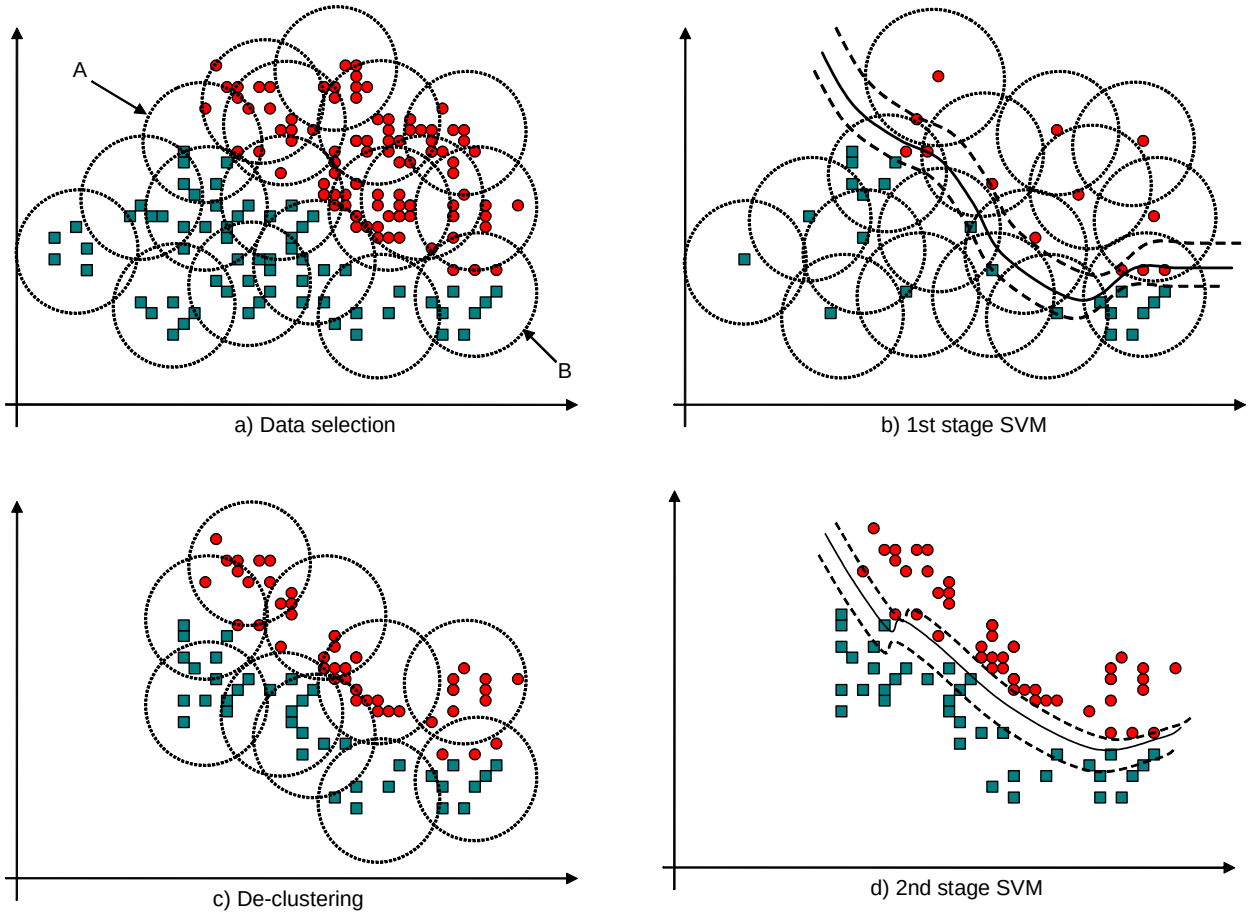
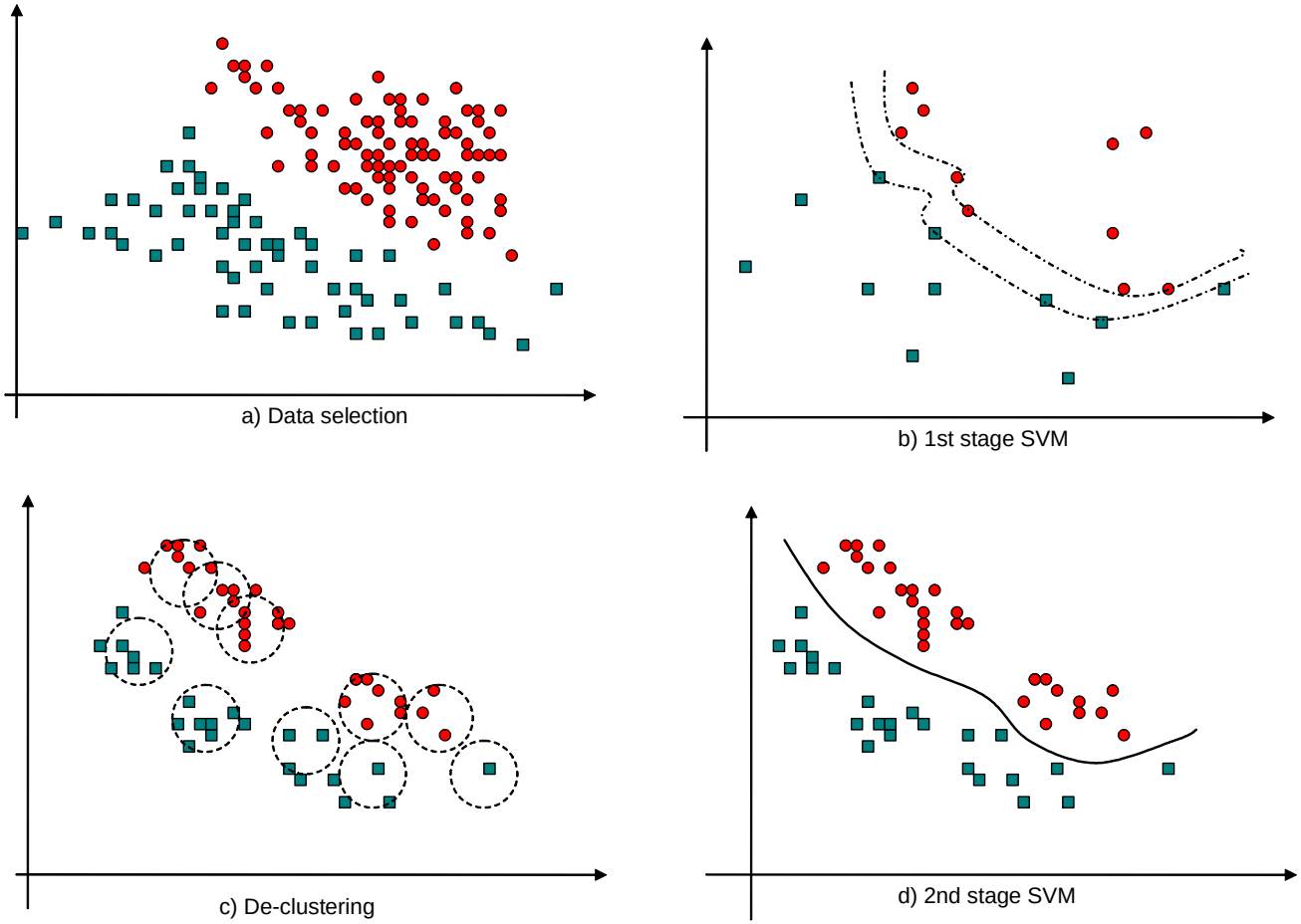


Figure:

RS-SVM^2

Two-stage classification via random selection



Example 1

synthetic data

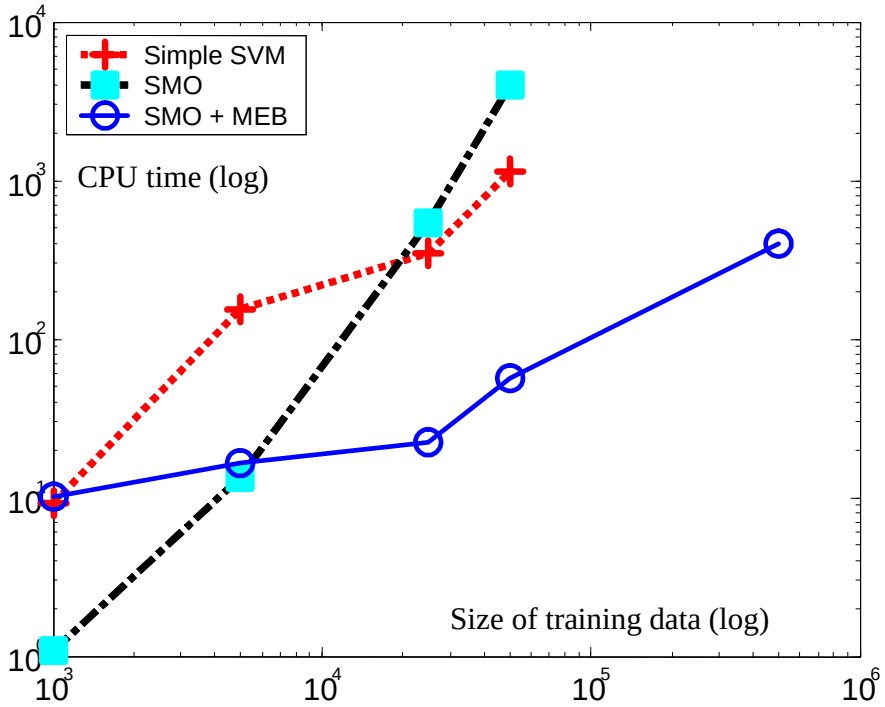
$500,000$ data were generated randomly in the range of $(0, 40)$ with two dimensions, i.e., $X_i = [x_{i,1}, x_{i,2}]$. The output (label) is decided as follows:

$$y_i = \begin{cases} +1 & \text{if } WX_i + b > th \\ -1 & \text{otherwise} \end{cases} \tag{3}$$

where $W = [1.2, 2.3]^T$, $b = 10$, $th = 95$.

Example 1

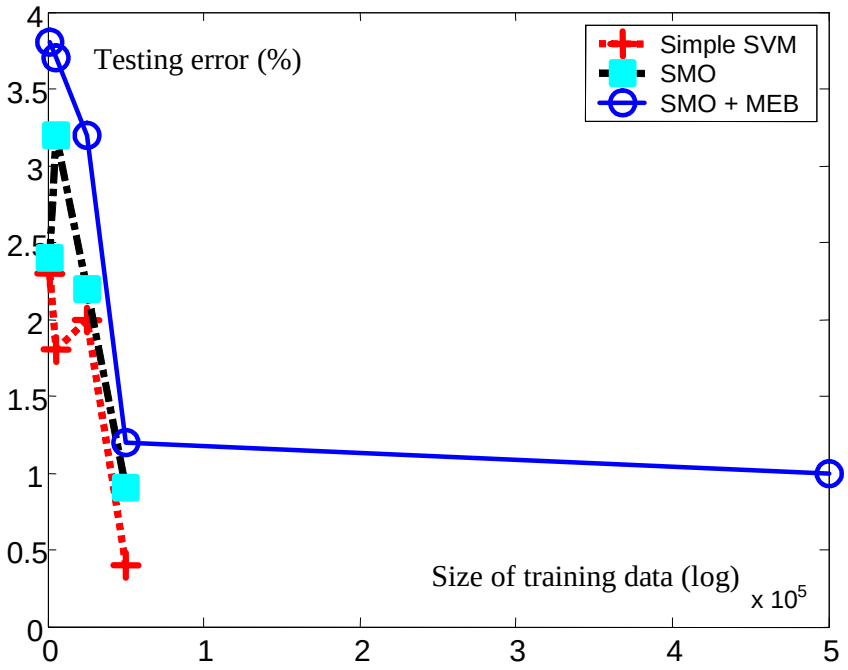
Synthetic data set



Example 1: running time *vs* training data size

Example 1

Synthetic data set



Example 1: testing accuracy *vs* training data size

Example 2

IJCNN 2001 benchmark

The data set is available at http://www.geocities.com/ijcnn/nnc_ijcnn01.pdf and http://www.csie.ntu.edu.tw/~cjlin/libsvm. There are 49990 training data points and 91701 testing data points, each record has 22 attributes. The sizes of the training data we used are $1,000$, $5,000$, $12,500$, $25,000$, $37,500$ and $49,990$.

Example 2

IJCNN 2001 benchmark

MEB-SVM^2

#	t	Acc	l	#MC	TrD1	SV1	TrD2	SV2
1000	22.34	93.8	350	39	388	125	199	51
5000	31.28	93.8	400	84	483	128	467	115
12500	38.41	95.2	450	105	554	105	733	160
25000	59.18	94.1	500	147	646	143	1342	228
37500	196.57	96.0	1000	179	1178	254	1399	267
49990	462.39	97.9	2000	201	2200	748	1728	295

RS-SVM^2

#	t	Acc	l	SV1	TrD2	SV2
1000	4.53	92.7	350	86	473	175
5000	8.73	94.1	400	109	541	180
12500	13.31	94.7	450	127	673	193
25000	25.98	94.9	500	118	712	287
37500	45.30	95.3	1000	122	1693	358
49990	78.08	97.7	2000	185	2370	430

Example 2

IJCNN 2001 benchmark

	MEB-SVM^2		RS-SVM^2.		Simple SVM		LIBSVM	
#	t	Acc	t	Acc	t	Acc	t	Acc
1000	22.3	93.8	4.5	92.7	2.1	94.3	0.5	93.1
5000	31.2	93.8	8.7	94.1	4.6	96.3	7.5	96.3
12500	38.4	95.2	13.3	94.7	182.6	96.1	47.6	98.2
25000	59.1	94.1	25.9	94.9	3823.0	97.2	177.0	98.6
37500	196.0	96.0	45.3	95.3	10872.0	97.7	394.0	98.8
49990	462.0	97.9	78.0	97.7	20491.0	98.2	730.0	98.8

Table 2. Training time and accuracy comparison between different algorithms on IJCNN 2001 data set

Example 3

RNA sequence data set

The RNA data set is available at *http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=1570369#top* from Supplementary Material (additional file 7). The data set consists of 23605 data points, each record has 8 attributes with continuous values between 0 to 1. The dataset contains 3919 ncRNAs and 19686 negative sequences. We used sizes 500, $1,000$, $2,500$, $5,000$, $10,000$ and $23,605$ in our experiments.

Example 3

RNA sequence data set

MEB-SVM^2

#	t	Acc	l	SV1	RD	SV2
500	4.7	85.3	350	87	397	168
1000	5.9	86.2	400	108	463	162
2500	15.5	86.3	450	124	529	209
5000	26.5	86.7	500	149	656	227
10000	69.3	86.9	650	199	862	282
23605	174.5	88.5	1500	278	1307	416

RS-SVM^2

#	t	Acc	l	SV1	RD	SV2
500	4.1	85.3	350	88	421	172
1000	4.4	85.7	400	97	453	153
2500	11.2	86.5	450	132	581	221
5000	15.8	86.1	500	146	637	211
10000	30.2	86.5	650	187	875	278
23605	65.7	88.3	1500	257	1275	381

Example 3

RNA sequence data set

	MEB-SVM^2		RS-SVM^2		LIBSVM		Simple SVM	
#	t	Acc	t	Acc	t	Acc	t	Acc
500	4.7	85.3	4.1	85.3	0.4	86.0	2.8	86.7
1000	5.9	86.2	4.4	85.7	0.7	87.2	8.2	87.1
2500	15.5	86.3	11.2	86.5	3.1	87.4	561.0	88.1
5000	26.5	86.7	15.7	86.1	12.5	87.6	—-	—-
10000	69.2	87.9	30.2	86.5	48.3	88.2	—-	—-
23605	174.0	88.2	65.7	88.3	298.0	88.6	—-	—-

Table 4. Training time and accuracy comparison between different algorithms on RNA sequence data set

Conclusion

We solve the trade-off problem between SVM classification accuracy and training time for large data sets, a two-stage SVM classification approach is proposed.

1. Our two stage classification approach is convenient for large data sets. But, not good for small data sets since data reduction might affect a lot on the accuracy.
2. Generally our approach can have almost the same accuracy as other SVM classifiers when the data set is large, while its training time is super shorter.
3. Random selection is faster than MEB and other clustering based data selection because it does not partition data, but it restricts that the original data set should be relatively uniform.
4. Two stage SVM classifier via MEB clustering may be the best method for general use comparing with other SVMs including two stage SVM via random selection.

Discussion

The accuracy decrease is caused by the lost of support vectors, several possible solutions are considered.

1. Increasing cluster number may increase the training data for the first stage SVM classification, so more support vectors may be obtained.
2. The relations between the clustering and the support vectors play an important role for classification accuracy. It may be solved from the point of data density.
3. Since clustering is unsupervised, some useful information (support vectors) of original data set may be lost. New clustering approaches which use label information may improve the accuracy. For example, the random selection of this paper is carried on in the two classes (labels ± 1) independently. This kind of method may be extended furthermore to more general semi-supervied clustering.

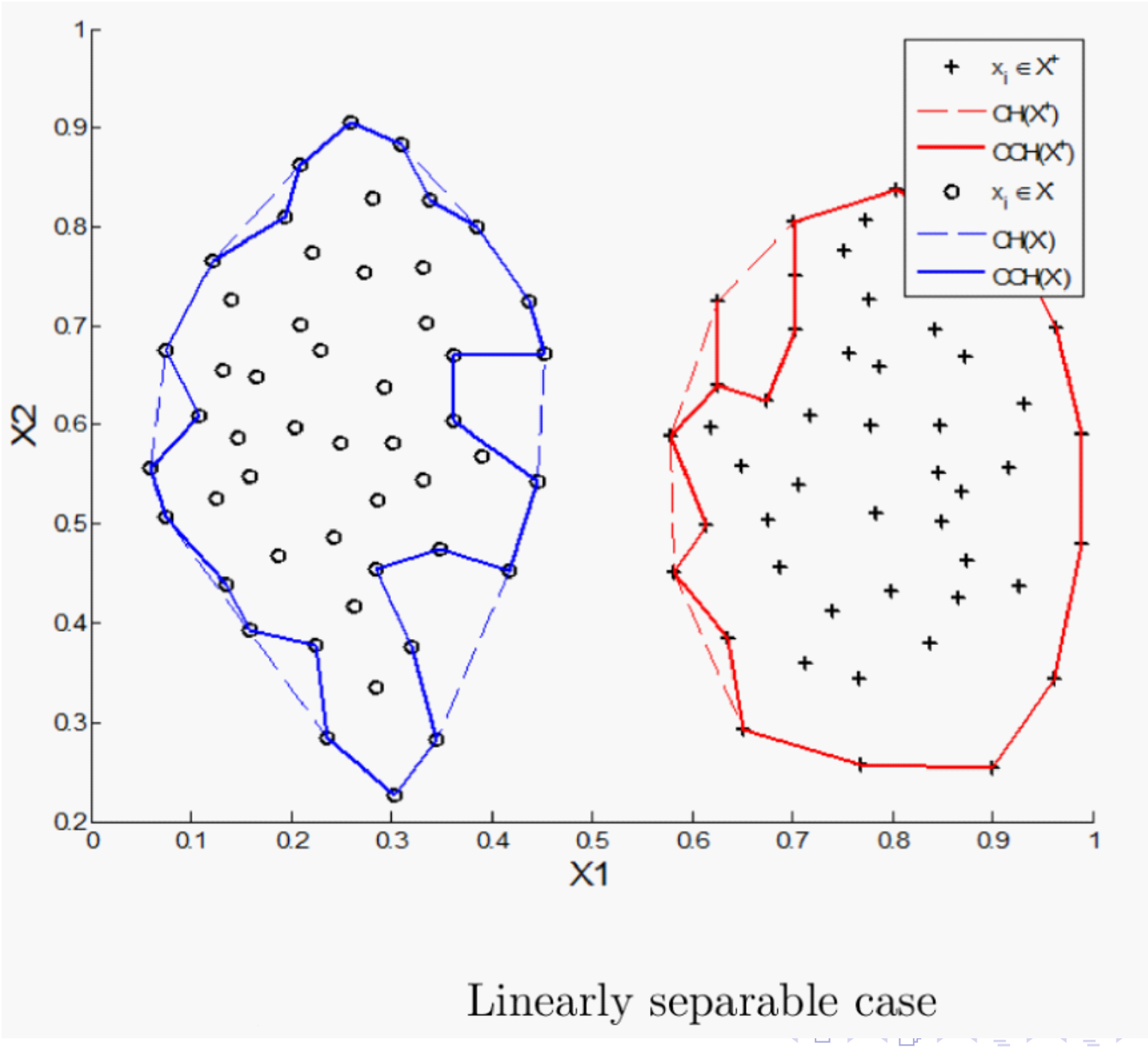
Other data reduction techniques for SVM Training

Main idea:

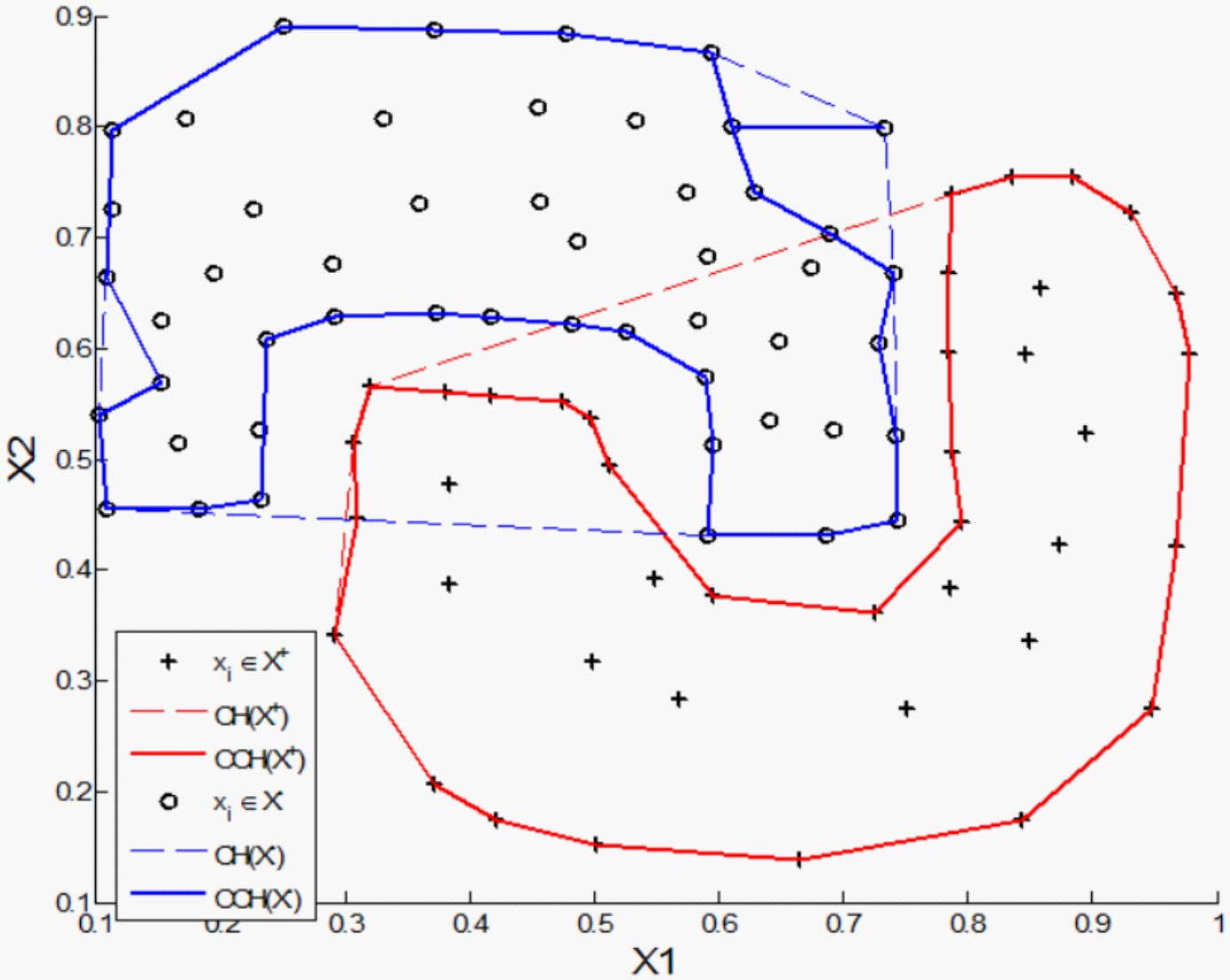
- Remove instances that are impossible to be SVs.
- The optimal separating hyperplane of SVM depends completely on instances located closest to the separation boundary.
- The number of SV is small compared with the size of entire data set.
- The same separating hyperplane is obtained if a SVM is trained with the whole training set or it is trained using only the SV.

Proposed methods

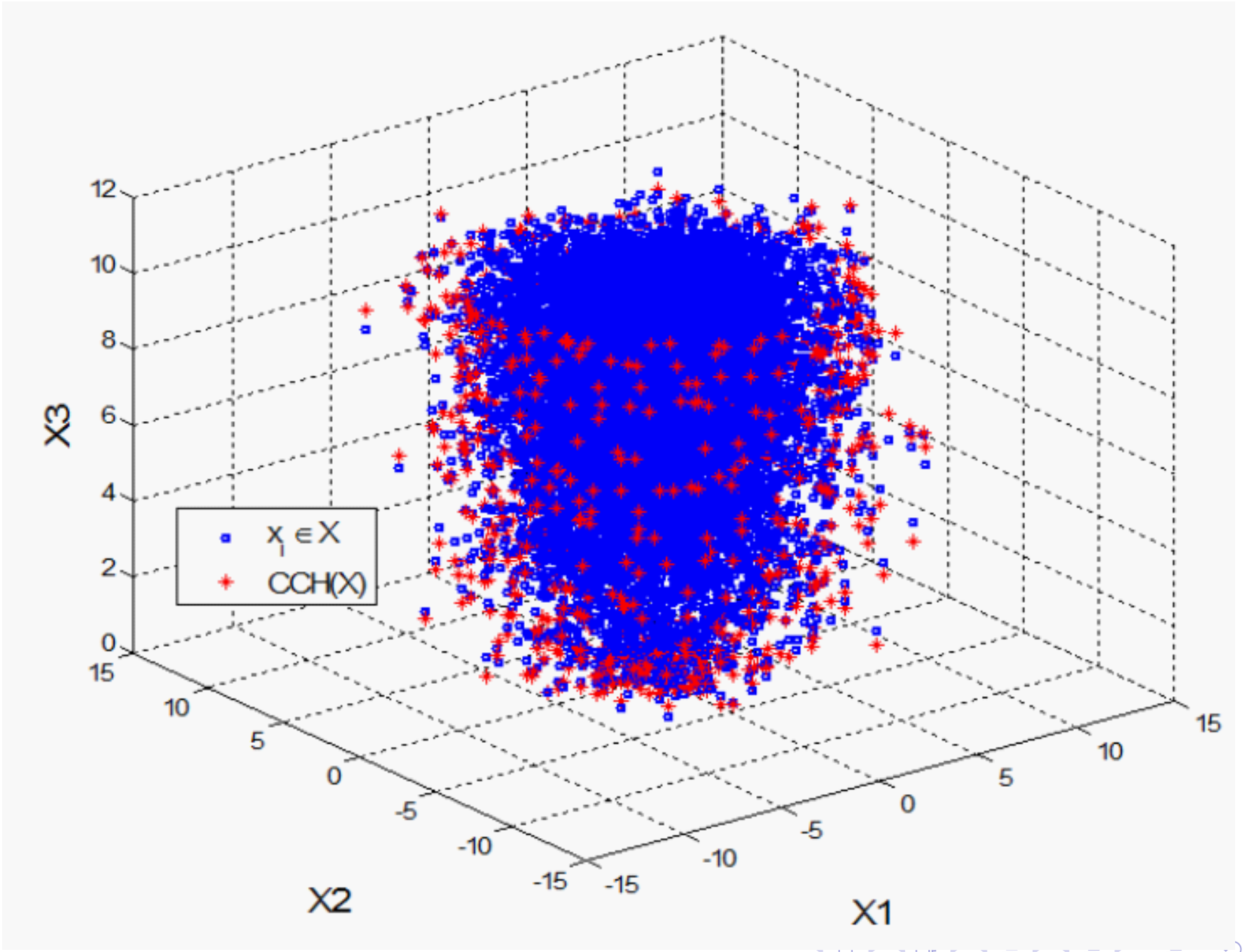
- CCH: Convex-concave hull.
- DTF: Decision tree and Fisher's linear discriminant.
- DTDRS: Decision Tree Directed Random Sampling



Linearly separable case



Linearly inseparable case


X3
X2
X1
$x_i \in X$
CCH(X)

Related publications (1)

1. Jair Cervantes, Xiaoou Li, Wen Yu, "Support vector machine classification for large data sets via minimum enclosing ball clustering", *Neurocomputing*, Volume(4-6): 611-619, 2008
2. Wen Yu, Xiaoou Li, "On-line fuzzy modeling via clustering and support vector machines", *Information Sciences*, 178: 4264-4279, 2008.
3. Xiaoou Li, Jair Cervantes, Wen Yu, Fast Classification for Large Data Sets via Random Selection Clustering and Support Vector Machines, *Intelligent Data Analysis*, 16(6): 897-914, 2012
4. Xiaoou Li, Wen Yu, Xiaoli Li, On-line Modeling via Fuzzy Support Vector Machines and Neural Networks, *Journal of Intelligent and Fuzzy Systems*, 24(3): 665-675, 2013
5. Jair Cervantes, Xiaoou Li, Wen Yu, Imbalanced Data Classification via Support Vector Machines and Genetic Algorithms, *Connection Science*, 26(4): 335-348, 2014

Related publications (2)

6 Asdrúbal López Chau, Xiaoou Li, Wen Yu, Large Datasets Classification Using Convex-Concave Hull and Support Vector Machine, *Soft Computing*, 17: 793-804, 2013

7 Asdrúbal López Chau, Xiaoou Li, Wen Yu, Convex and Concave Hulls for Classification with Support Vector Machine, *Neurocomputing*, 122: 198-209, 2013

8 Asdrúbal López, Xiaoou Li, Wen Yu, Support Vector Machine Classification for Large Data Sets Using Decision Tree and Fisher Linear Discriminant, *Future Generation Computer Systems*, 36(1): 57-65, 2014

9 Suresh Thenozhi, Wen Yu, Asdrúbal López Chau, and Xiaoou Li, "Structural Health Monitoring of Tall Buildings with Numerical Integrator and Convex-Concave Hull Classification," Mathematical Problems in Engineering, vol. 2012, Article ID 212369, 15 pages, 2012.

10 Xiaoou Li and Wen Yu, Fast support vector machine classification for large data sets, *International Journal of Computational Intelligence*