# Enhancing SVM with visualization

Thanh-Nghi DO, François POULET

***dtnghi@lri.fr, poulet@esiea-ouest.fr***

# Outline of talk

- introduction
- SVM algorithms
  - proximal SVM (PSVM)
  - reduced SVM (RSVM)
- enhancing SVM with visualization
  - multiple views
  - cooperation Viz-RSVM for classification
  - pre-processing for large data sets
  - interpreting SVM results
- numerical test results
- conclusion and future work

- **introduction**
- SVM algorithms
  - proximal SVM (PSVM)
  - reduced SVM (RSVM)
- enhancing SVM with visualization
  - multiple views
  - cooperation Viz-RSVM for classification
  - pre-processing for large data sets
  - interpreting SVM results
- numerical test results
- conclusion and future work

# Introduction

- data in the world increase rapidly
  - WalMart: 20M transactions/day
  - Google: 70M researches/day
  - AT&T: 275M calls/day

- data mining
  - necessary
  - discover knowledge in large databases
  - algorithms: decision tree, clustering, association rules, ***support vector machines*** (Vapnik, 1995), etc.

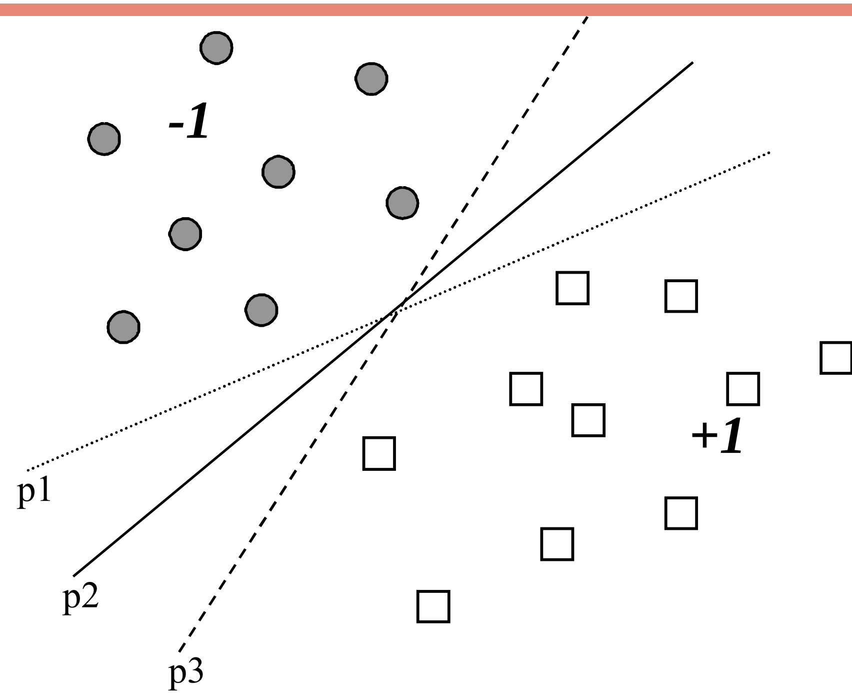
# Introduction

- support vector machines
  - optimal surface for separating data into two classes
  - classification, regression, novelty detection.
  - successful applications: face identification, text categorization, bioinformatics, etc
  - but SVM results: incomprehensible

- contribution
  - enhancing SVM with visualization
  - classification task
  - interpreting SVM results

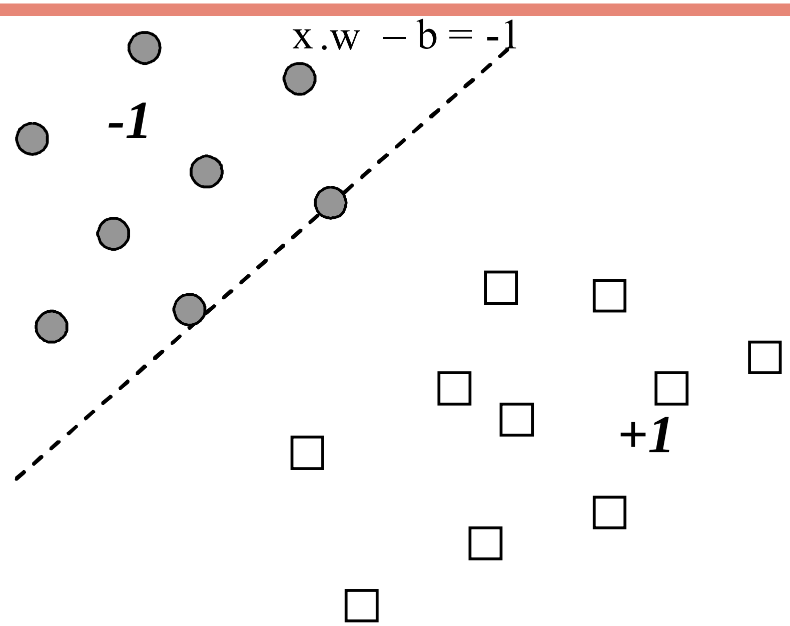
- introduction
- **SVM algorithms**
  - proximal SVM (PSVM)
  - reduced SVM (RSVM)
- enhancing SVM with visualization
  - multiple views
  - cooperation Viz-RSVM for classification
  - pre-processing for large data sets
  - interpreting SVM results
- numerical test results
- conclusion and future work

# Support Vector Machines

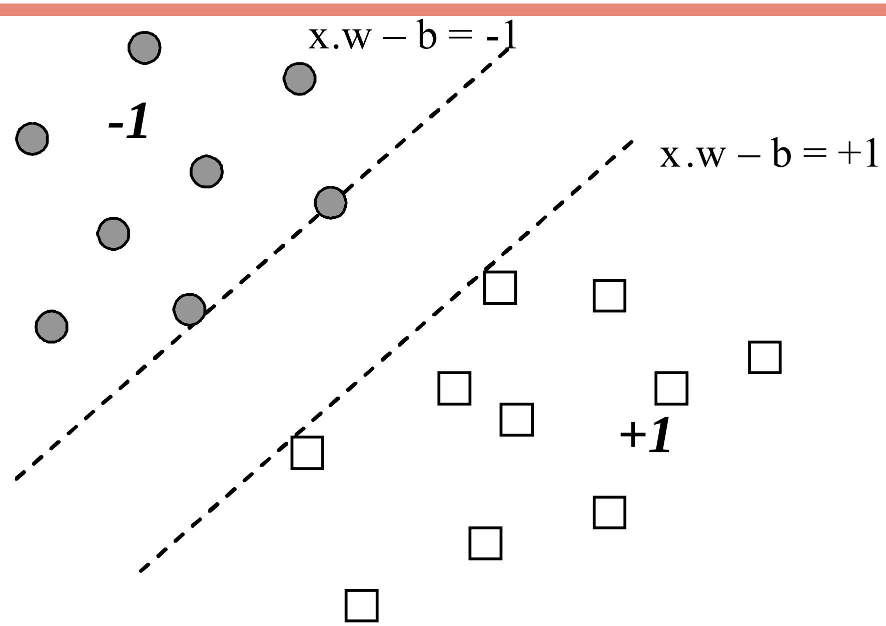



- how to find optimal separating plane (w, b)
  - classify m points $x_1, x_2, \ldots, x_n$ in n-dimensions into 2 classes $\pm1$

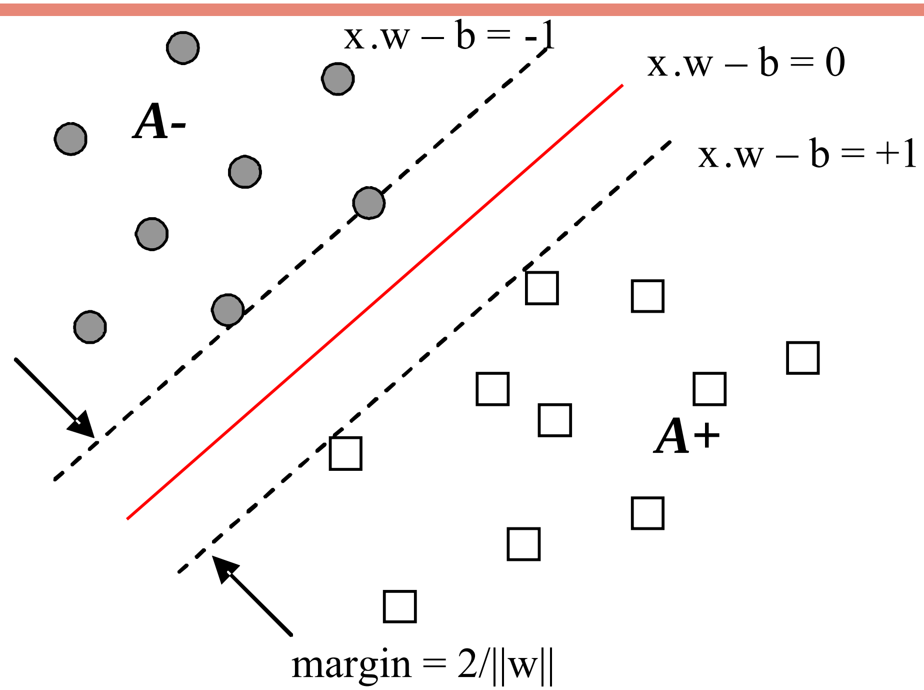
# Support Vector Machines



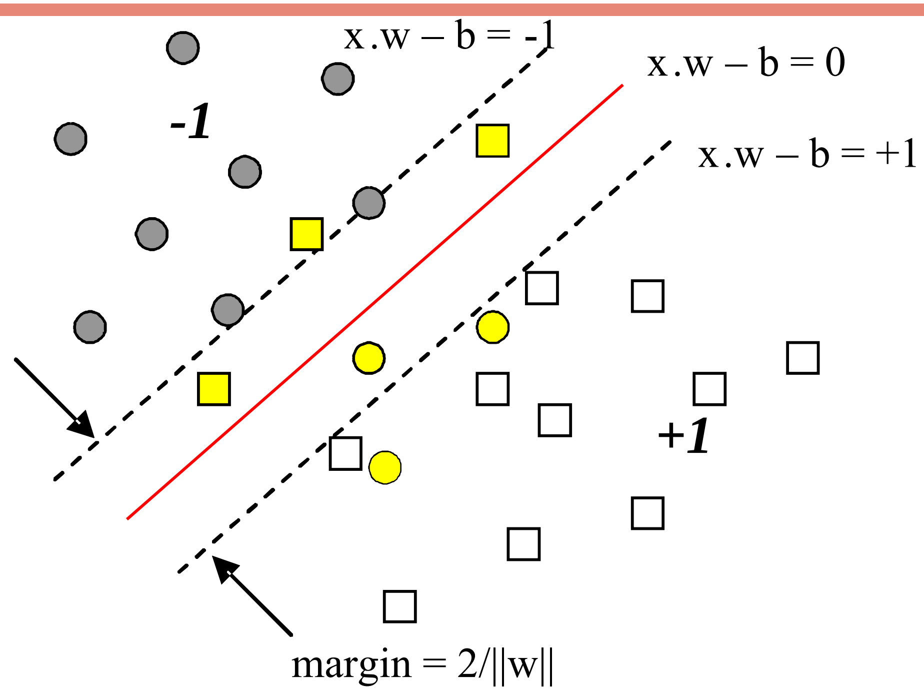
# Support Vector Machines



# Support Vector Machines



# Support Vector Machines




- soft margin: errors are allowed

# SVM algorithm

- maximizing margin + minimizing errors
  - be accomplished through quadratic program
  - solution: w, b
- classify new x: sign (x.w – b)

# Proximal SVM (Fung & Mangasarian, 2001)

- PSVM changes:
  - inequality constraints to equalities
  - maximizing margin to min (1/2) $||w, b||^2$
  - minimizing least squares 2-norm errors
- PSVM = solution of linear system (n+1) variables (w,b)
  - instead of quadratic program solution
  - very fast to train
- non linear PSVM
  - kernel matrix K with mxm size (m: nb. data points)
  - high cost (memory, time)

# RSVM (Lee & Mangasarian, 2000)

- reducing problem size
  - using random data points (size r) from whole dataset
  - creating kernel matrix K with mxr size (r << m)
  - drastically reducing cost
  - comparable solution

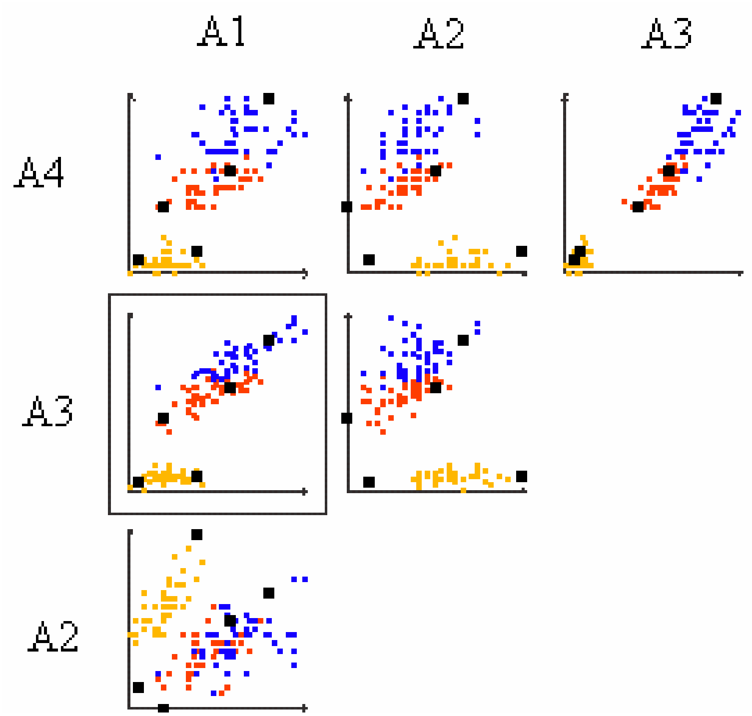
- using random subset r as support vectors : not clear
- no idea for creating kernel matrix

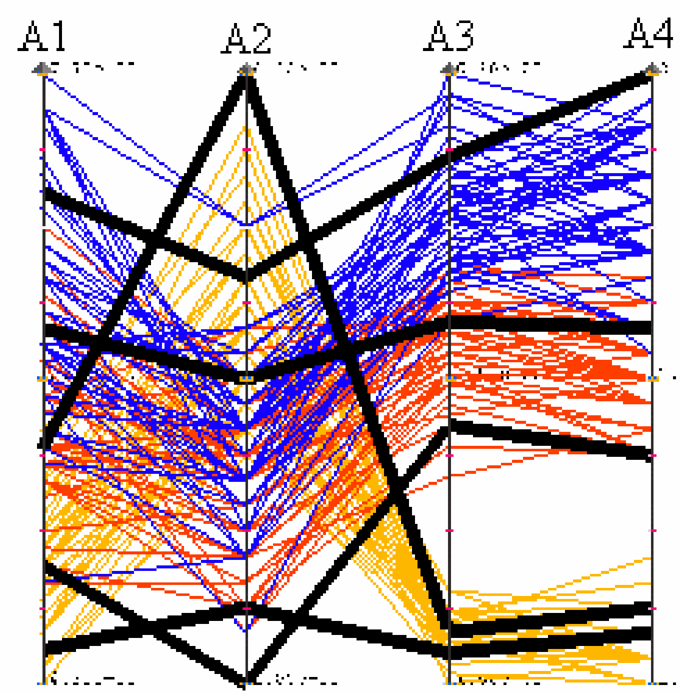
- introduction
- SVM algorithms
  - proximal SVM (PSVM)
  - reduced SVM (RSVM)
- **<u>enhancing SVM with visualization</u>**
  - multiple views
  - cooperation Viz-RSVM for classification
  - pre-processing for large data sets
  - interpreting SVM results
- numerical test results
- conclusion and future work

# Cooperation visualization-RSVM

- visualizing dataset
- select support vectors for RSVM
  - brushing data points close to separating boundary
- overview of data found by appropriate visualization
  - some idea for choosing & tuning kernel
- interpreting SVM results with multi-view
  - improving model comprehensibility
  - detecting interesting dimensions

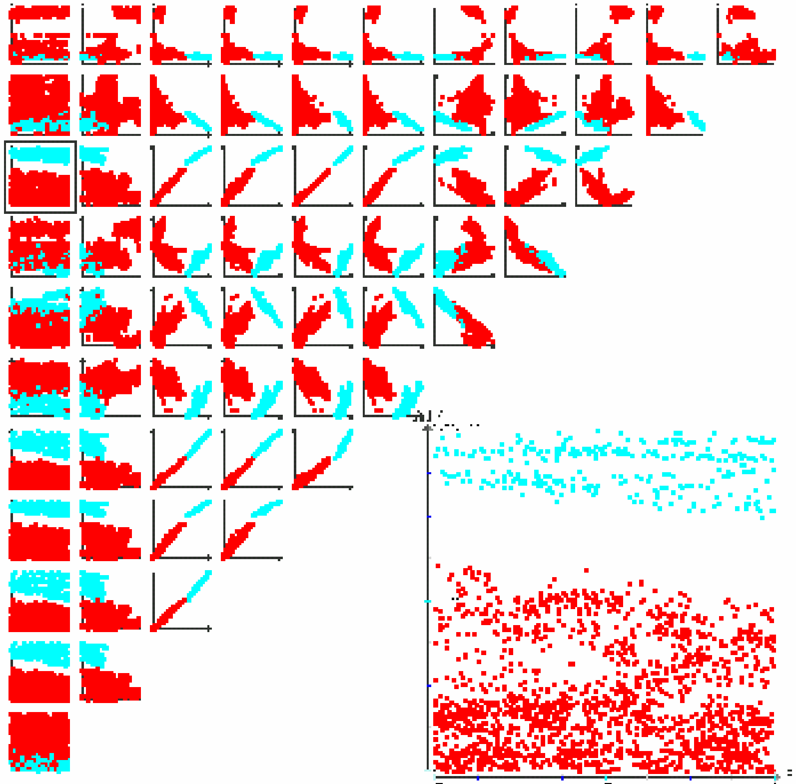
# Multi-view, linking, brushing

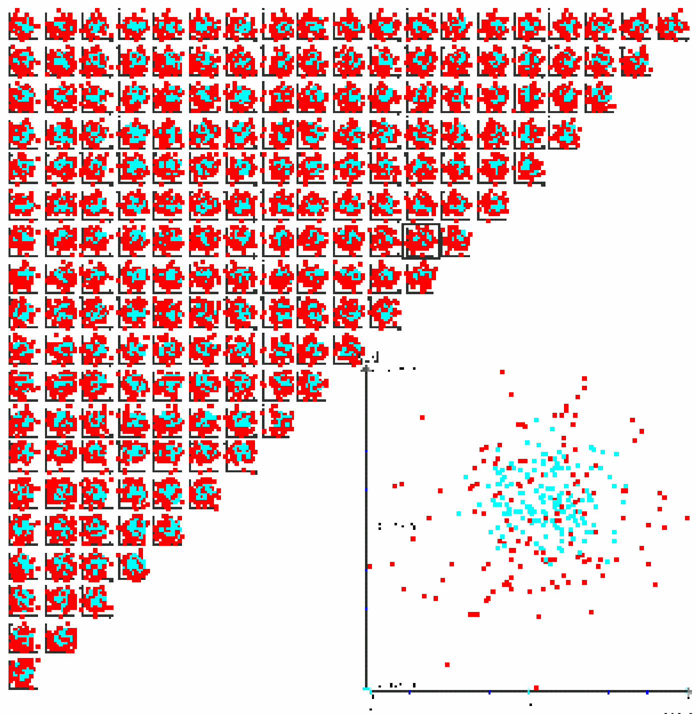



- 2D scatter-plot matrices
  - 2D projections
  - color = class
- parallel coordinates
  - point => polyline
  - color = class

# Getting overview of data

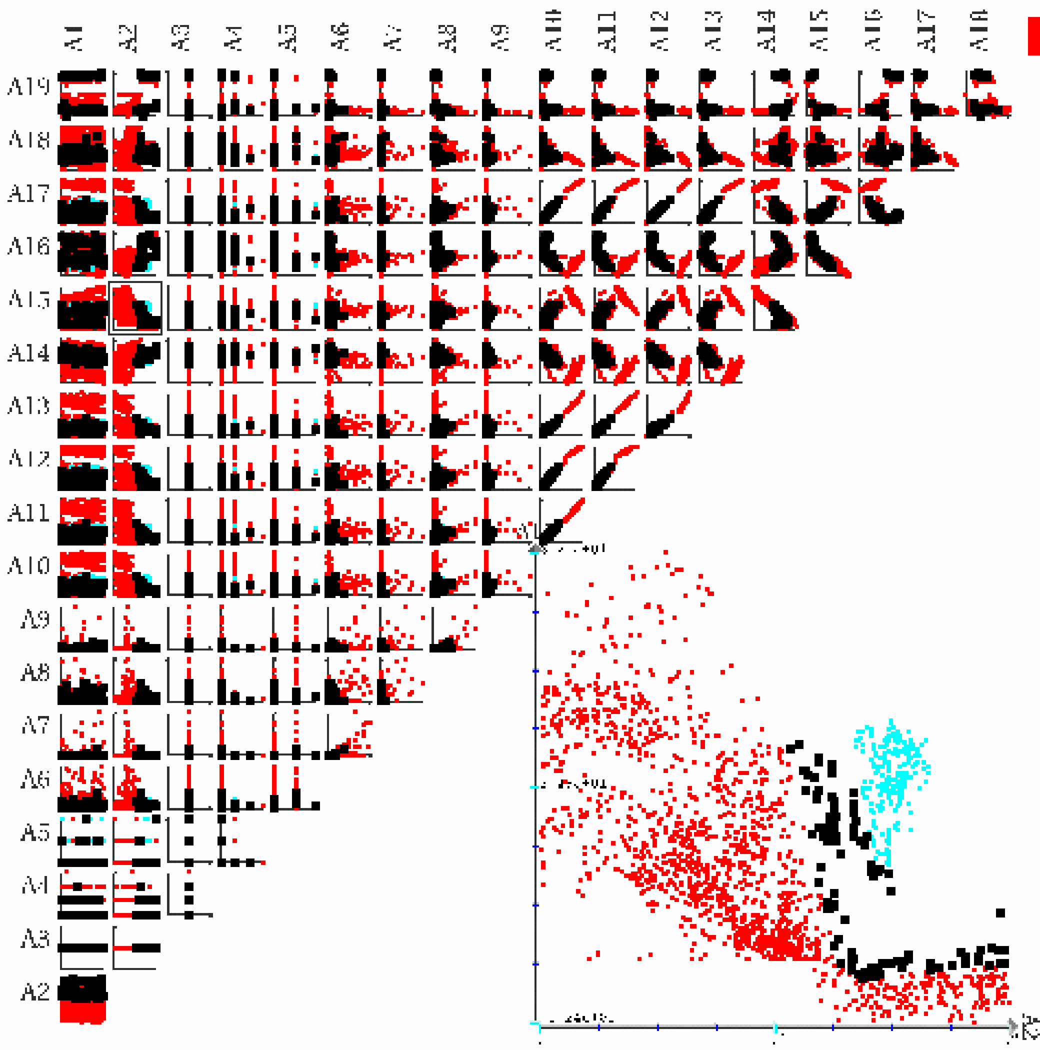




- view of dataset1
  - linear separating
  - linear classification task
- view of dataset2
  - non linear separating
  - non linear classification task

# Interactive brushing support vectors

(separating 6 “cyan” against all “red” of Segment dataset, 2310 points, 19 dims, 7 classes)




- select support vectors
  - data points close to separating boundary
  - non linear separation
  - kernel function : RBF
  - $K(x,y) = \exp(-\text{gamma}||x-y||^2)$
  - increasing gamma if boundary close
  - creating kernel matrix K witl gamma = 0.0001 because frontier not complex
  - RSVM : 100 % accuracy

# Pre-processing for mining large datasets

- large number points
  - creating clusters with SOM or K-Means
  - sampling subset from clusters
- large number dimensions
  - feature selection by 1-norm SVM
  - getting dimension subset

# Classification on large nb. points
(separating 2 against all of Forest dataset, 581012 points, 54 dims, 7 classes)

- SVMTorch has used 100,000 training, 50,000 testing
  - 2 days 5 hours with 83.24 % accuracy (on Athlon 1.2 GHz)
- we have used 500,000 training, 50,000 testing
  - LibSVM has not finished learning task after several days
  - created 200 clusters with SOM
  - sampled 10,000 points from clusters
  - Viz-RSVM has finished learning task in 8 hours (on P-4, 2.4 GHz)
  - with 84.32 % accuracy

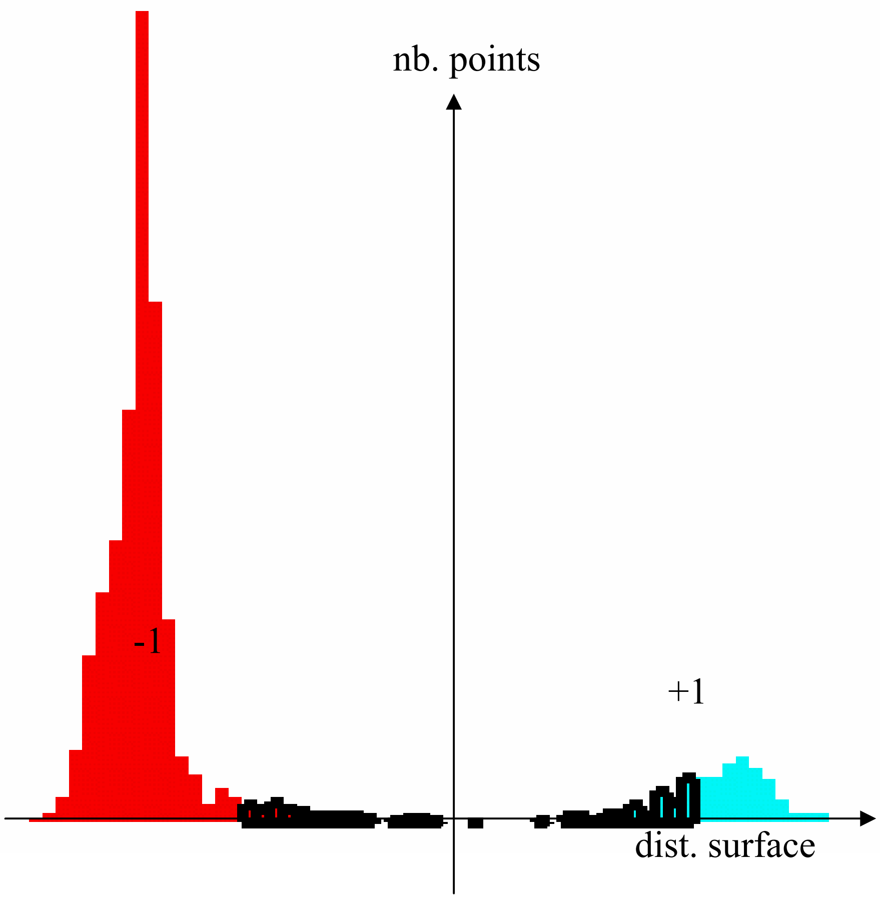
# Classification on large nb. dims
## (bio-medical datasets)

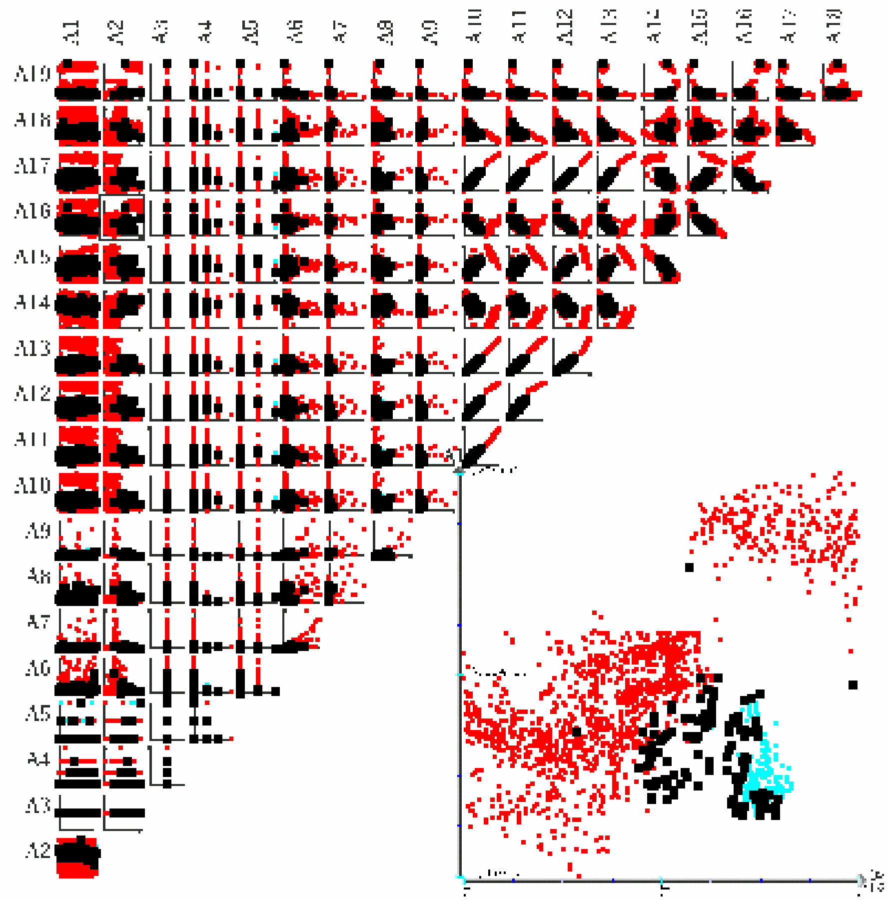
| | Accuracy of +1 | | Accuracy of -1 | | Accuracy | |
|---|---|---|---|---|---|---|
| | Feature selection | No selection | Feature Selection | No selection | Feature selection | No Selection |
| AML-ALL Leukemia | **100 %** 5-D | 95 % 7129-D | 85.71 % 5-D | **92.86 %** 7129-D | **94.12 %** 5-D | **94.12 %** 7129-D |
| Breast Cancer | **91.67 %** 10-D | 83.33 % 24481-D | **57.14 %** 10-D | **57.14 %** 24481-D | **78.95 %** 10-D | 73.68 % 24481-D |
| Colon Tumor | **95.45 %** 19-D | 86.36 % 2000-D | **97.5 %** 19-D | 92.5 % 2000-D | **96.77 %** 19-D | 90.32 % 2000-D |
| Lung Cancer | **100 %** 9-D | **100 %** 12533-D | 96.27 % 9-D | **98.51 %** 12533-D | 96.64 % 9-D | **98.66 %** 12533-D |
| Ovarian Cancer | **100 %** 13-D | **100 %** 15154-D | **100 %** 10-D | **100 %** 15154-D | **100 %** 13-D | **100 %** 15154-D |

- feature selection is best (except Lung Cancer data)
  - interesting dimensions
  - visualizing results

# Interpreting SVM results
## - visualizing separating boundary
## - detecting interesting dimensions

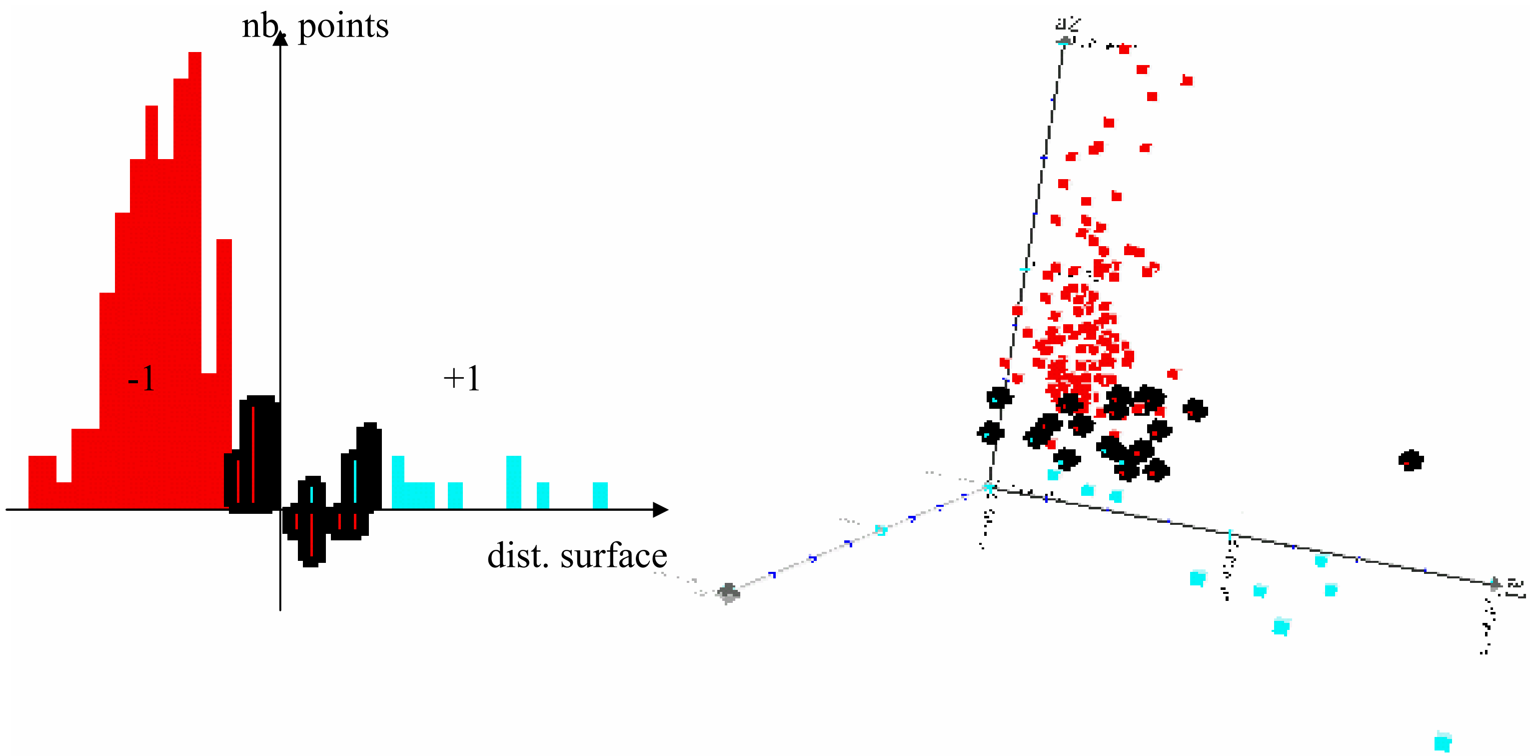




- data distribution
- 2D scatter-plot matrices

# Visualizing result
(Lung Cancer, 149 points, 12533 dims, 2 classes)




- feature selection by 1-norm SVM
  - need only 9 dims for classification
  - 3D view presents clearly separating boundary
  - 3 corresponding dimensions are interesting result

- introduction
- SVM algorithms
  - proximal SVM (PSVM)
  - reduced SVM (RSVM)
- enhancing SVM with visualization
  - multiple views
  - cooperation Viz-RSVM for classification
  - pre-processing for large data sets
  - interpreting SVM results
- **numerical test results**
- conclusion and future work

# Performances evaluation

| | Classes | Points | Attributes | Evaluation method |
|---|---|---|---|---|
| Bupa | 2 | 345 | 6 | 10-fold |
| Pima | 2 | 768 | 8 | 10-fold |
| Twonorm | 2 | 7400 | 20 | 300 trn – 7100 tst |
| Ringnorm | 2 | 7400 | 20 | 300 trn – 7100 tst |
| Segment | 7 | 2310 | 19 | 10-fold |
| Satimage | 6 | 6435 | 36 | 4435 trn – 2000 tst |
| Forest | 7 | 581012 | 54 | 500000 trn – 50000 tst |

- software program
  - our cooperative method Viz-RSVM
  - automatic SVM algorithm LibSVM

# Results

| | Viz-RSVM | LibSVM |
|---|---|---|
| Bupa | **76.18 %** | 73.62 % |
| Pima | **78.86 %** | 77.34 % |
| Twonorm | 97.28 % | **97.35 %** |
| Ringnorm | 97.15 % | **97.28 %** |
| Segment | 96.02 % | **97.10 %** |
| Satimage | 91.70 % | **92.05%** |
| Forest | **84.32 %** | N/A |

- Viz-RSVM gives good results compared with LibSVM
  - kernel construction is helped by human pattern recognition capacities + visualization
  - gaining insight into classification task with visualization
  - understanding results

- introduction
- SVM algorithms
  - proximal SVM (PSVM)
  - reduced SVM (RSVM)
- enhancing SVM with visualization
  - multiple views
  - cooperation Viz-RSVM for classification
  - pre-processing for large data sets
  - interpreting SVM results
- numerical test results
- **<u>conclusion and future work</u>**

# Conclusion

- enhancing SVM with visualization method
  - selecting support vectors
  - providing some idea for kernel construction
  - interpreting results
  - looking inside SVM “black box”
  - comparable solution with automatic algorithm
  - pre-processing for large datasets

# Future work

- better use of user domain knowledge for mining task
- extract useful rules for interpreting SVM results
- mining large datasets via symbolic objects

Thanks !