Genomic, Proteomic and Transcriptomic Lab
High Performance Computing and Networking Institute
National Research Council, Italy

Mathematical Models of Supervised Learning and their Application to Medical Diagnosis

Mario Rosario Guarracino

January 9, 2007

Consiglio Nazionale delle Ricerche

Acknowledgements

- prof. Franco Giannessi – U. of Pisa,
- prof. Panos Pardalos – CAO UFL,
- Onur Seref – CAO UFL,
- Claudio Cifarelli – HP.

Agenda

- ▶Mathematical models of supervised learning
- ▶Purpose of incremental learning
- ▶Subset selection algorithm
- ▶Initial points selection
- ▶Accuracy results
- ▶Conclusion and future work

Introduction

- *Supervised learning* refers to the capability of a system to learn from examples (*training set*).
- The trained system is able to provide an answer (*output*) for each new question (*input*).
- *Supervised* means the desired output for the training set is provided by an external teacher.
- *Binary classification* is among the most successful methods for supervised learning.

Applications

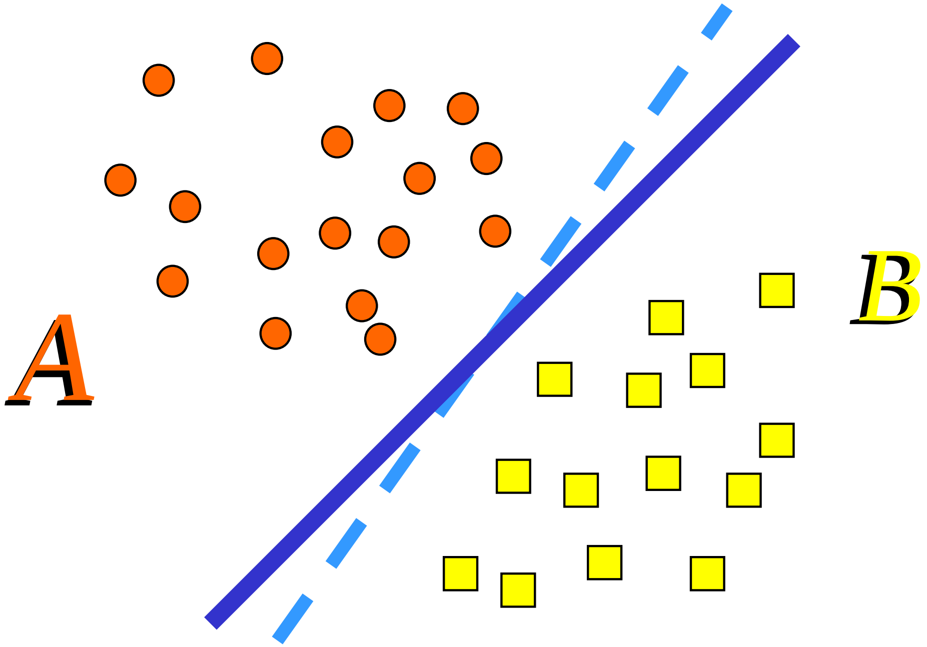
- Many applications in biology and medicine:
  - Tissues that are prone to cancer can be detected with high accuracy.
  - Identification of new genes or isoforms of gene expressions in large datasets.
  - New DNA sequences or proteins can be tracked down to their origins.
  - Analysis and reduction of data spatiality and principal characteristics for drug design.

Problem characteristics

- Data produced in biomedical application will exponentially increase in the next years.
- Gene expression data contain tens of thousand characteristics.
- In genomic/proteomic application, data are often updated, which poses problems to the training step.
- Current classification methods can over-fit the problem, providing models that do not generalize well.

Linear discriminant planes

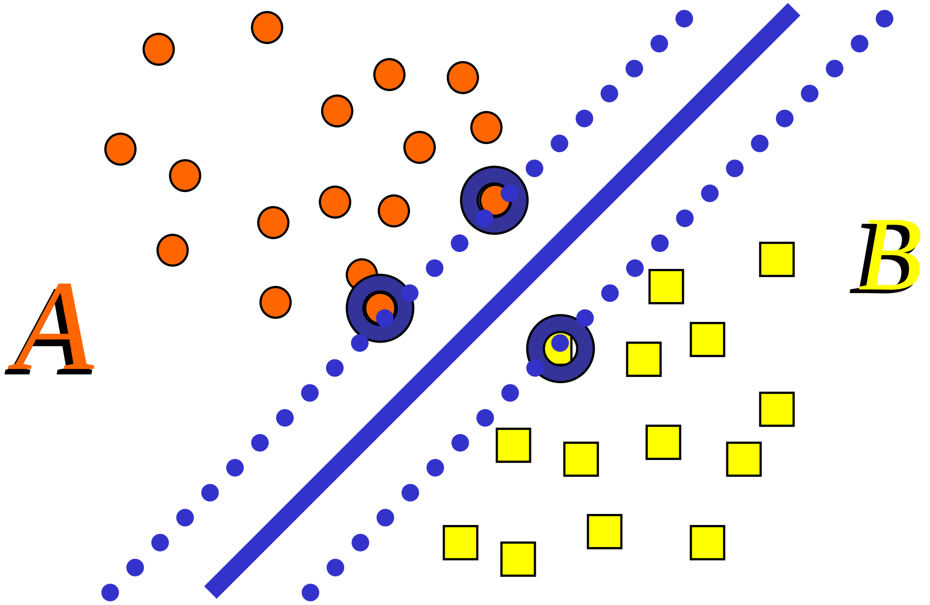
- Consider a binary classification task with points in two linearly separable sets.
  - There exists a plane that classifies all points in the two sets



- There are infinitely many planes that correctly classify the training data.

SVM classification

- A different approach, yielding the same solution, is to maximize the margin between *support planes*
  - Support planes leave all points of a class on one side



$$\min_{a} \frac{1}{2}\|w\|^2$$

$$s.t.$$

$$Aw + b \geq e$$

$$Bw + b < -e$$

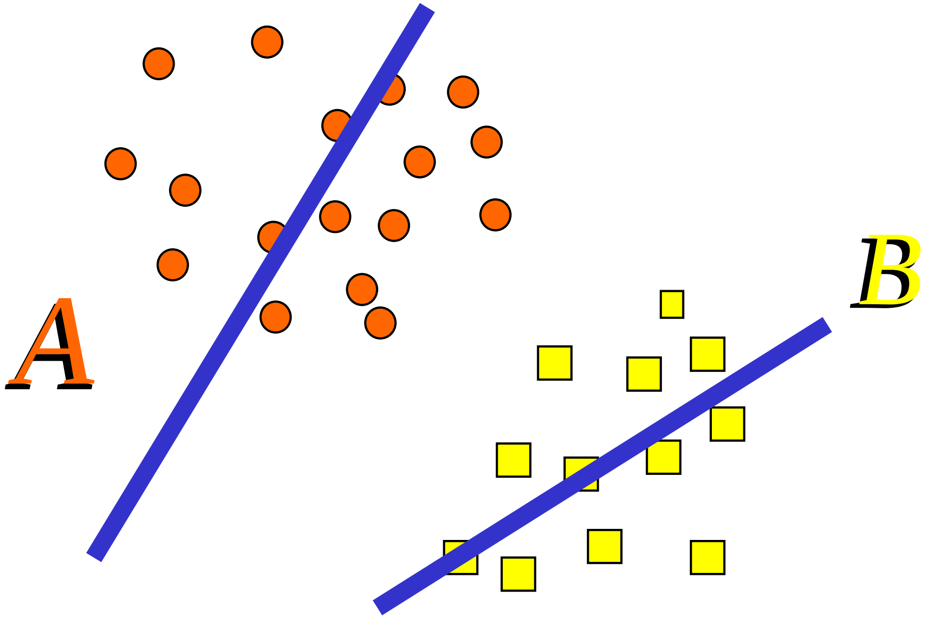
- Support planes are pushed apart until they “bump” into a small set of data points (*support vectors*).

SVM classification

- Support Vector Machines are the state of the art for the existing classification methods.
- Their robustness is due to the strong fundamentals of statistical learning theory.
- The training relies on optimization of a quadratic convex cost function, for which many methods are available.
  - Available software includes SVM-Lite and LIBSVM.
- These techniques can be extended to the nonlinear discrimination, embedding the data in a nonlinear space using *kernel functions*.

A different religion

- Binary classification problem can be formulated as a generalized eigenvalue problem (GEPSVM).
- Find $x'w_1=\gamma_1$ the closer to A and the farther from B:



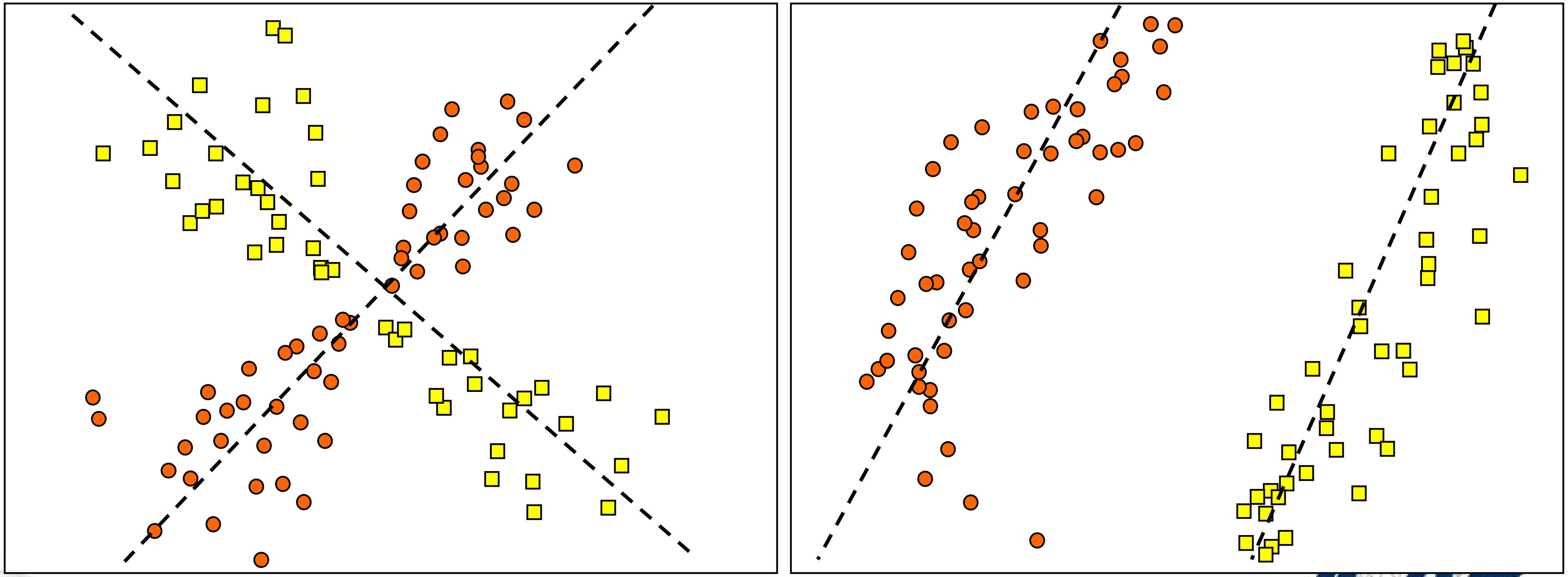
$$\min_{w,\gamma\neq 0} \frac{\|Aw - e\gamma\|^2}{\|Bw - e\gamma\|^2}$$

O. Mangasarian *et al.*, (2006) *IEEE Trans. PAMI*

ReGEC technique

Let $[w_1\ \gamma_1]$ and $[w_m\ \gamma_m]$ be eigenvectors associated to min and max eigenvalues of $Gx=\lambda Hx$:

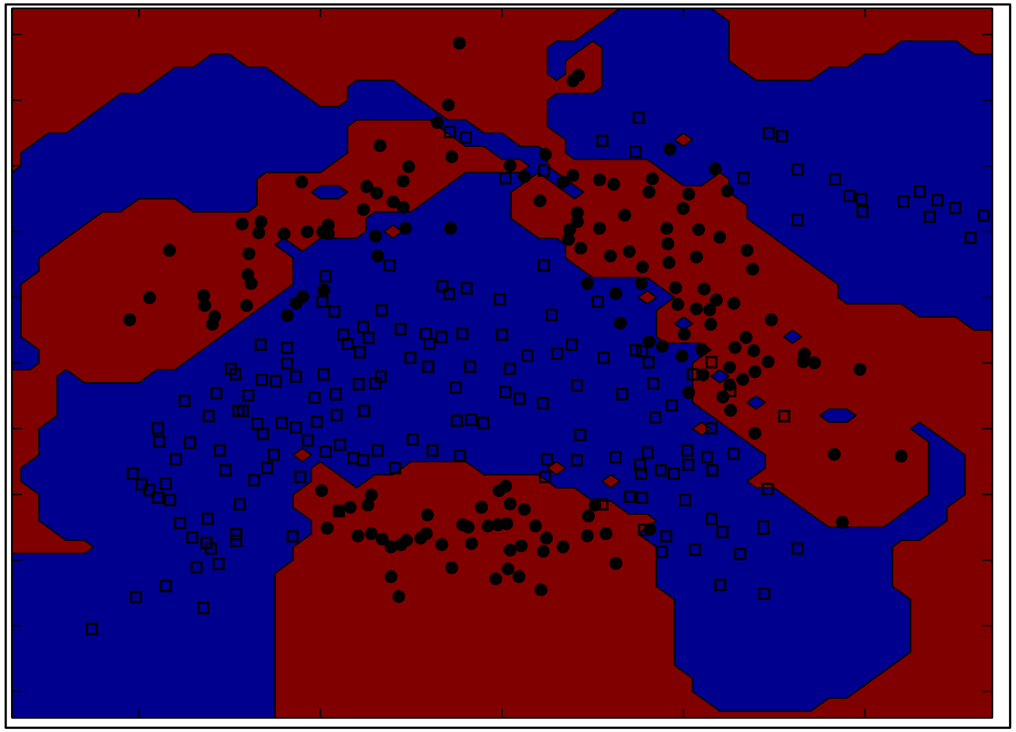
- $a \in A \Leftrightarrow$ closer to $x'w_1 - \gamma_1 = 0$ than to $x'w_m - \gamma_m = 0$,
- $b \in B \Leftrightarrow$ closer to $x'w_m - \gamma_m = 0$ than to $x'w_1 - \gamma_1 = 0$.



M.R. Guarracino *et al.*, (2007) *OMS.*

Nonlinear classification

- When classes cannot be linearly separated, nonlinear discrimination is needed.



- Classification surfaces can be very tangled.
- This model accurately describes original data, but does not generalize to new data (*over-fitting*).

How to solve the problem?

Incremental classification

- A possible solution is to find a small and robust subset of the training set that provides comparable accuracy results.

- A smaller set of points:
  - reduces the probability of over-fitting the problem,
  - is computationally more efficient in predicting new points.

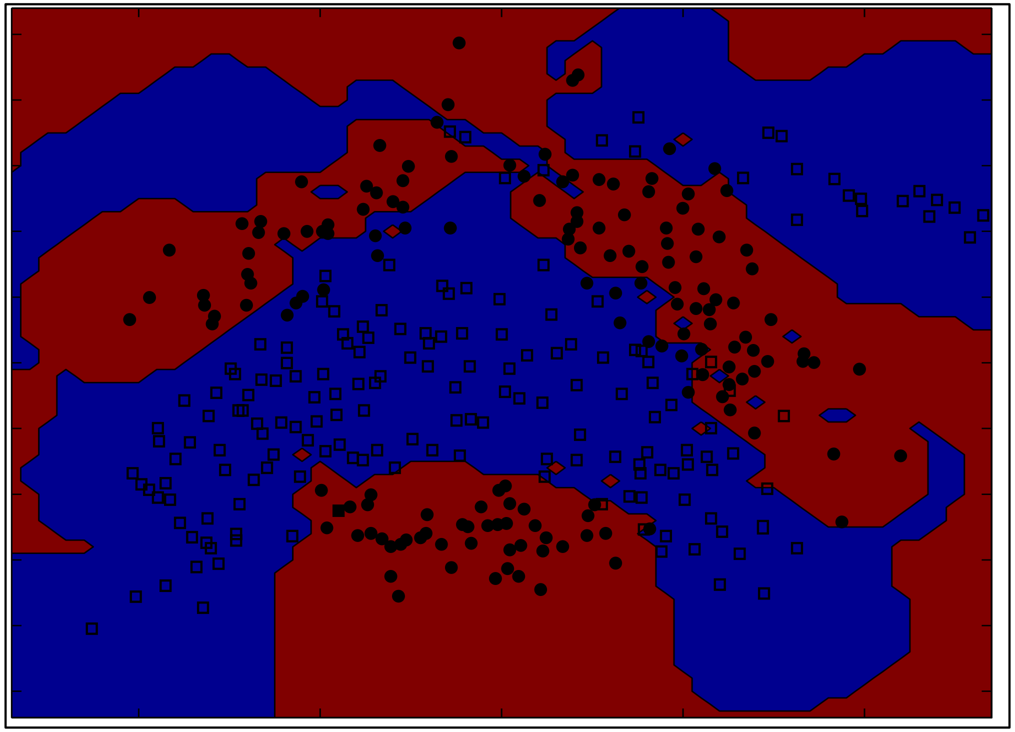
- As new points become available, the cost of retraining the algorithm decreases if the influence of the new points is only evaluated with respect to the small subset.

I-ReGEC: Incremental learning algorithm

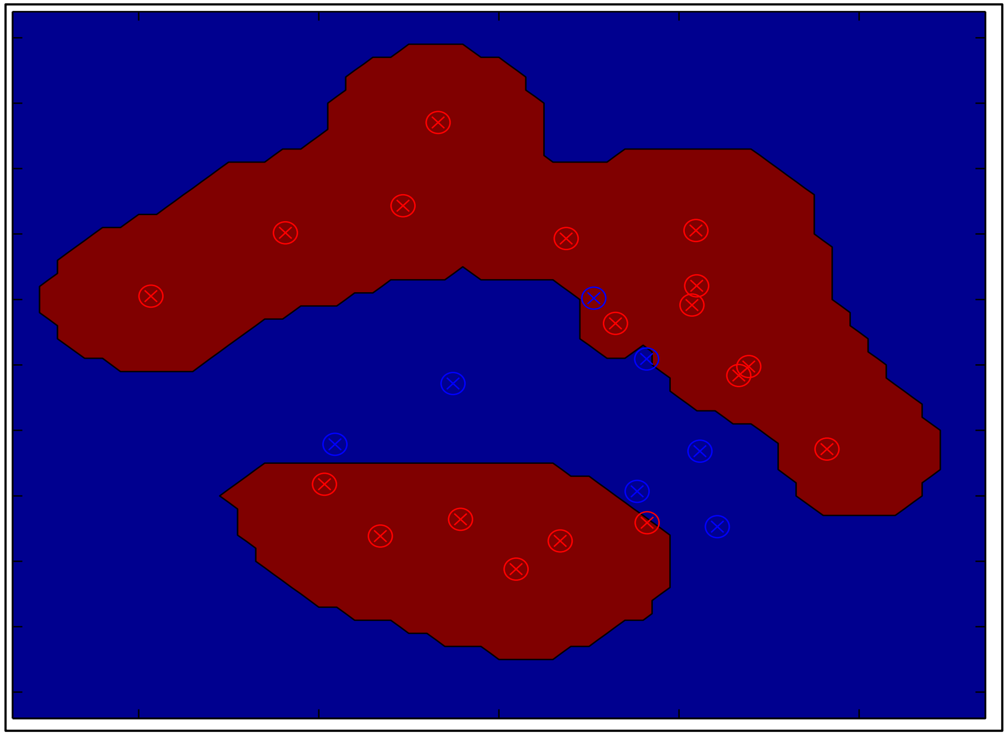
```
$\Gamma_0 = C \setminus C_0$
$\{M_0, Acc_0\} = Classify( C; C_0 )$
$k = 1$
**while** $|\Gamma_k| > 0$ **do**
    $x_k = x : \max_{x \in \{M_k \cap \Gamma_{k-1}\}} \{dist(x, P_{class(x)})\}$
    $\{M_k, Acc_k\} = Classify( C; \{C_{k-1} \cup \{x_k\}\} )$
    **if** $Acc_k > Acc_{k-1}$ **then**
        $C_k = C_{k-1} \cup \{x_k\}$
        $k = k + 1$
    **end if**
    $\Gamma_k = \Gamma_{k-1} \setminus \{x_k\}$
**end while**
```

I-ReGEC overfitting

ReGEC accuracy=84.44



I-ReGEC accuracy=85.49



- When ReGEC algorithm is trained on all points, surfaces are affected by noisy points (*left*).
- I-ReGEC achieves clearly defined boundaries, preserving accuracy (*right*).
  - **Less then 5% of points needed for training!**

Initial points selection

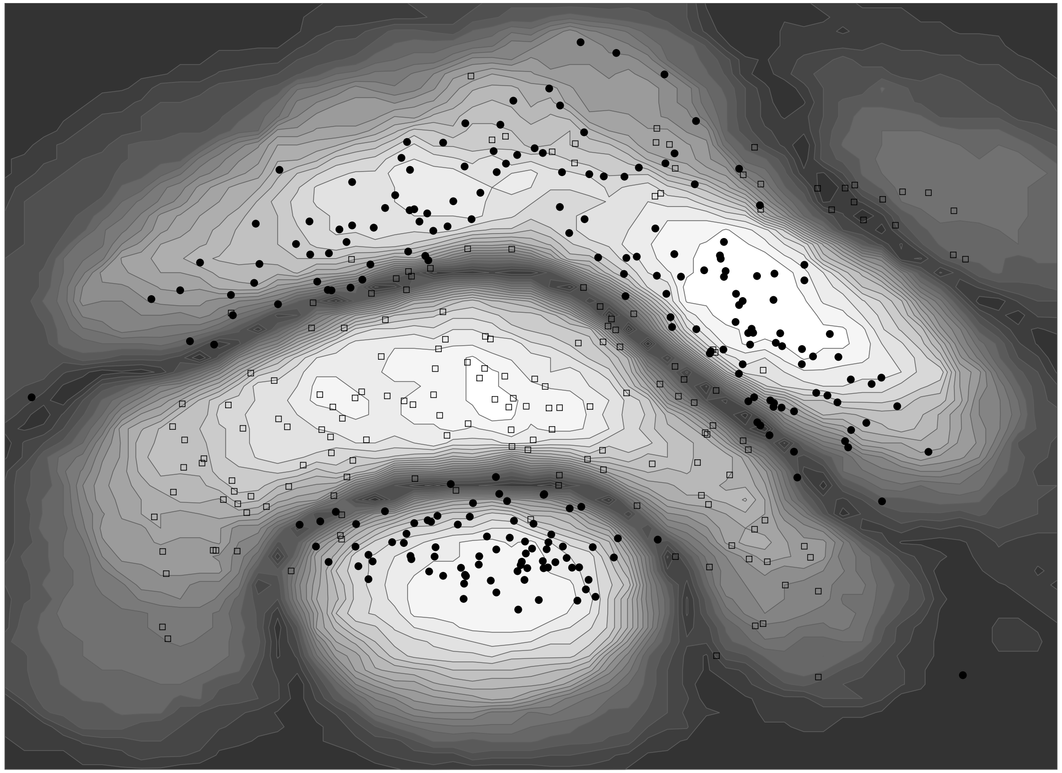
- Unsupervised clustering techniques can be adapted to select initial points.

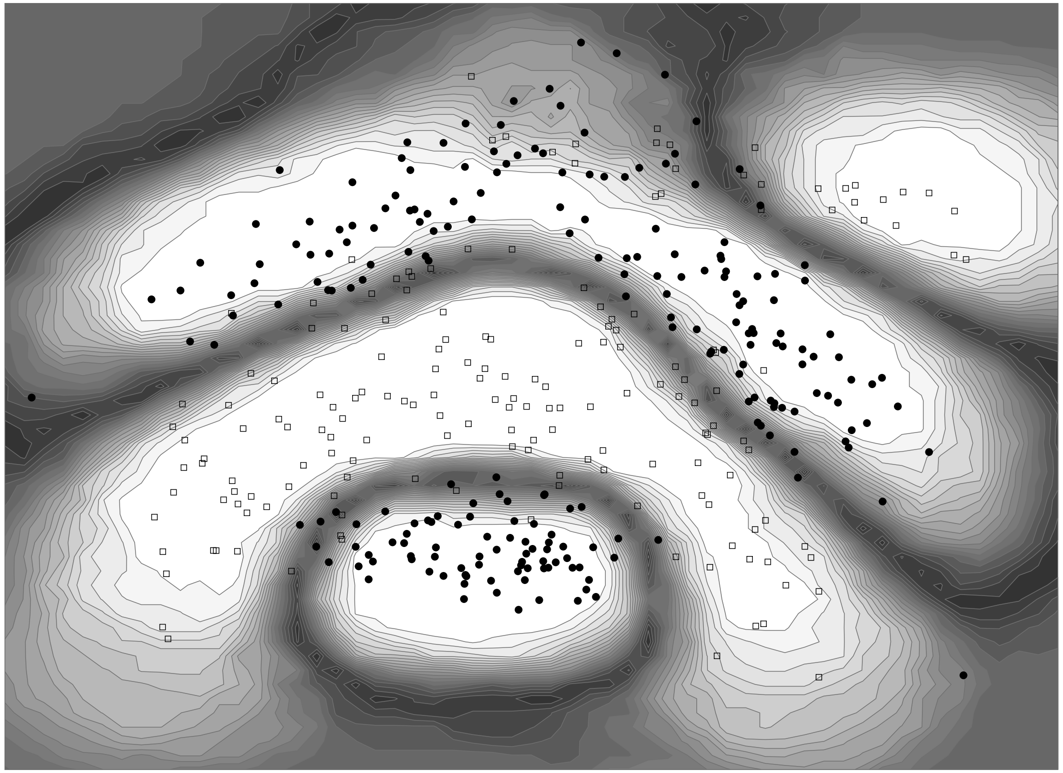
- We compare the classification obtained with k randomly selected starting points for each class, and k points determined by *k-means* method.

- Results show higher classification accuracy and a more consistent representation of the training set, when *k-means* method is used instead of random selection.

Initial points selection

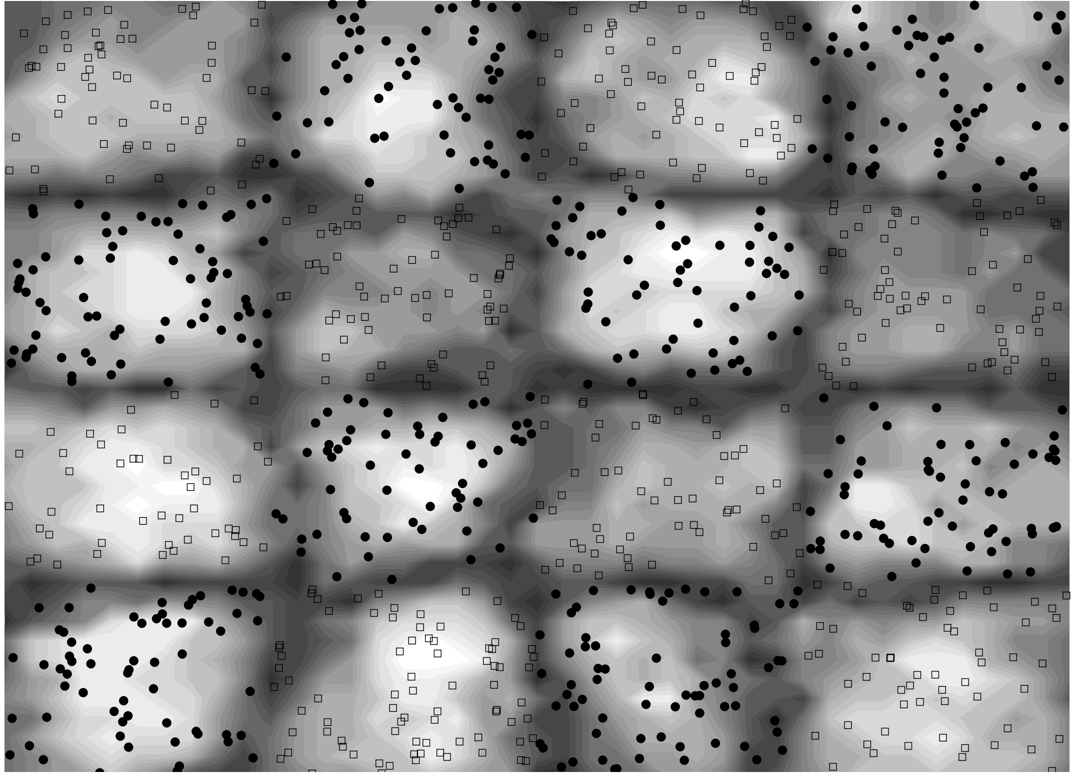
- Starting points C_i chosen:
  - randomly (top),
  - k-means (bottom).
- For each kernel produced by C_i, a set of evenly distributed points x is classified.
  - The procedure is repeated 100 times.
- Let $y_i \in \{1; -1\}$ be the classification based on C_i.
- $y = |\sum y_i|$ estimates the probability x is classified in one class.
  - random acc=84.5 std = 0.05
  - k-means acc=85.5 std = 0.01

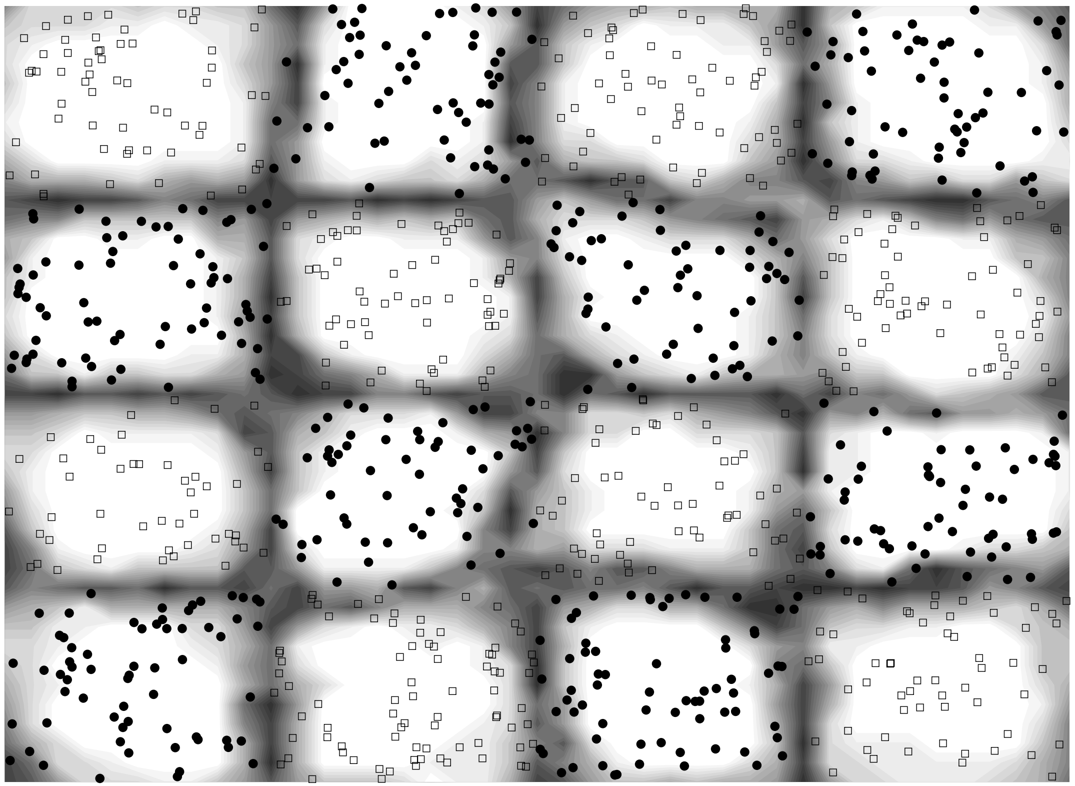




Initial points selection

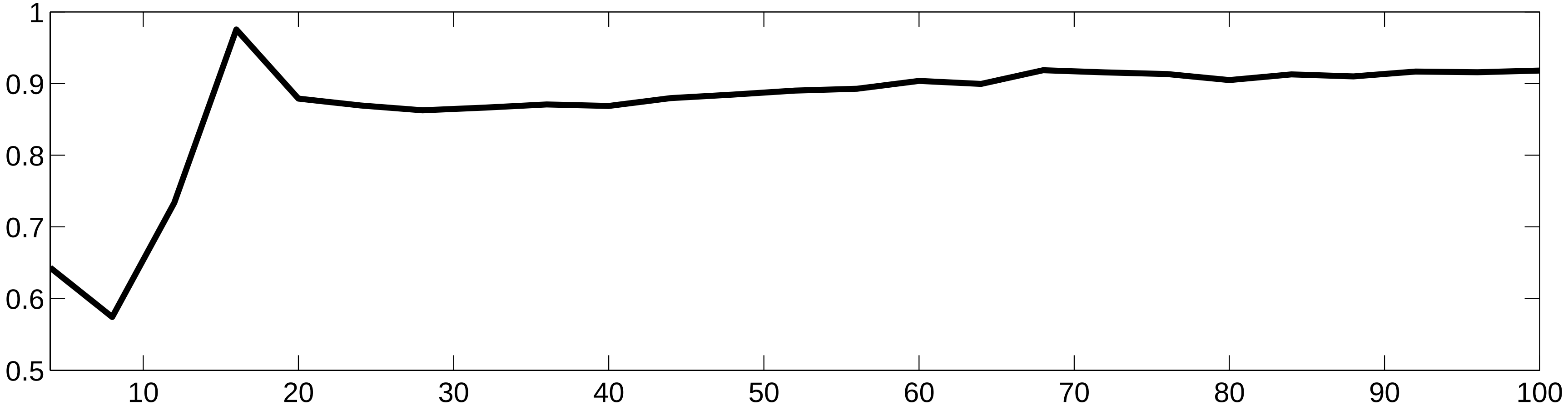
- Starting points C_i chosen:
  - randomly (top),
  - k-means (bottom).
- For each kernel produced by C_i, a set of evenly distributed points x is classified.
  - The procedure is repeated 100 times.
- Let $y_i \in \{1; -1\}$ be the classification based on C_i.
- $y = |\sum y_i|$ estimates the probability x is classified in one class.
  - random acc=72.1std = 1.45
  - k-means acc=97.6 std = 0.04





Initial point selection

- Effect of increasing initial points k with *k-means* on Chessboard dataset.

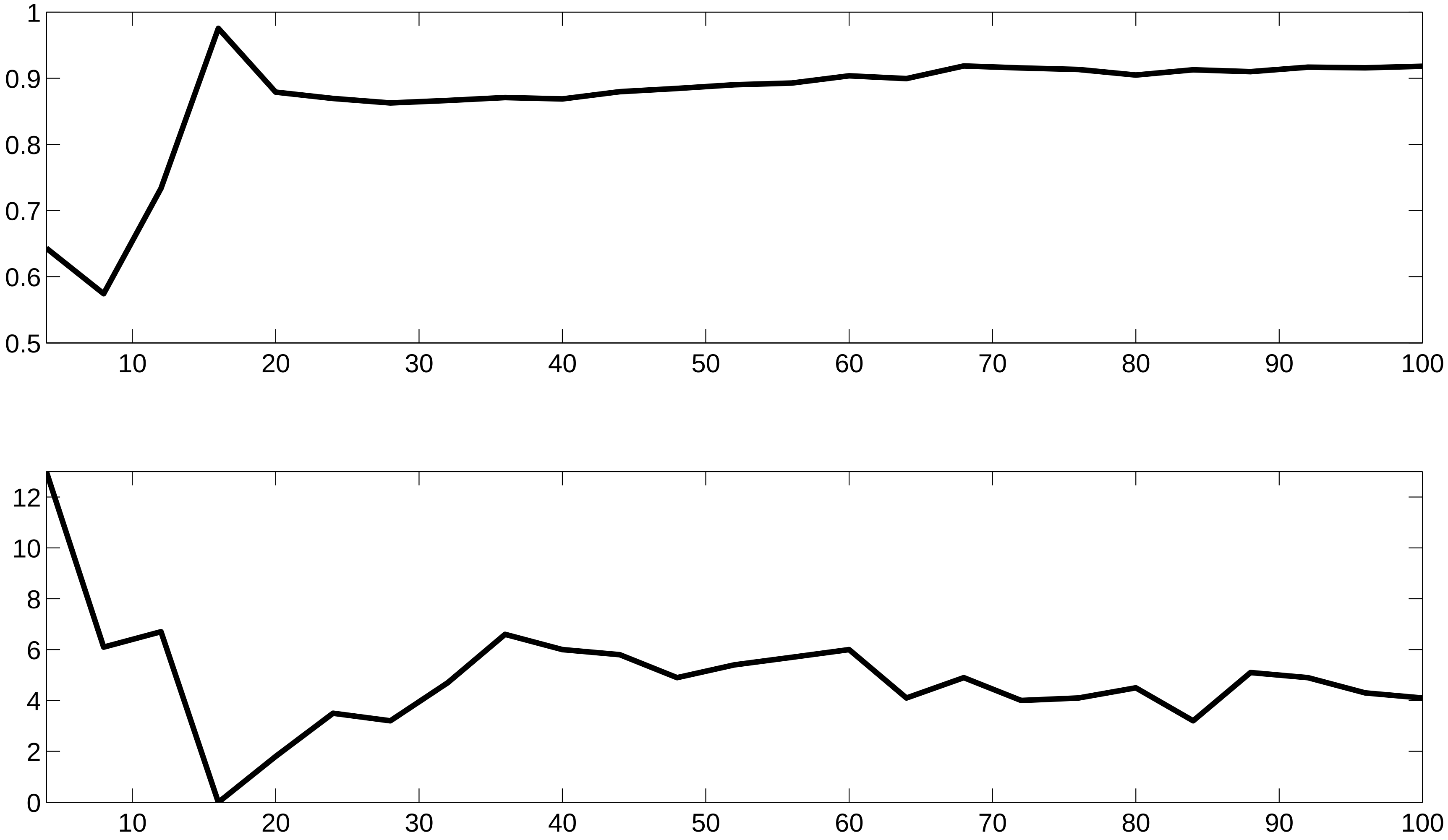


- The graph shows the classification accuracy versus the total number of initial points $2k$ from both classes.

- This result empirically shows that there is a minimum k, for which maximum accuracy is reached.

Initial point selection

- Bottom figure shows k vs. the number of additional points included in the incremental dataset.



Dataset reduction

- Experiments on real and synthetic datasets confirm training data reduction.

	I-ReGEC	
Dataset	chunk	% of train
Banana	15.7	3.92
German	29.09	4.15
Diabetis	16.63	3.55
Haberman	7.59	2.76
Bupa	15.28	4.92
Votes	25.9	6.62
WPBC	4.215	4.25
Thyroid	12.40	8.85
Flare-solar	9.67	1.45

Accuracy results

- Classification accuracy with incremental techniques well compare with standard methods

	ReGEC		I-ReGEC			SVM
Dataset	train	acc	chunk	k	acc	acc
Banana	400	84.44	15.70	5	85.49	89.15
German	700	70.26	29.09	8	73.5	75.66
Diabetis	468	74.56	16.63	5	74.13	76.21
Haberman	275	73.26	7.59	2	73.45	71.70
Bupa	310	59.03	15.28	4	63.94	69.90
Votes	391	95.09	25.90	10	93.41	95.60
WPBC	99	58.36	42.15	2	60.27	63.60
Thyroid	140	92.76	12.40	5	94.01	95.20
Flare-solar	666	58.23	9.67	3	65.11	65.80

Positive results

- Incremental learning, in conjunction with ReGEC, reduces training sets dimension.
- Accuracy results well compare with those obtained selecting all training points.
- Classification surfaces can be generalized.

Ongoing research

- Microarray technology can scan expression levels of tens of thousands of genes to classify patients in different groups.
- For example, it is possible to classify types of cancers with respect to the patterns of gene activity in the tumor cells.
- Standard methods fail to derive grouping of genes responsible of classification.

Examples of microarray analysis

- Breast cancer: *BRCA1* vs. *BRCA2* and sporadic mutations,
  - I. Hedenfalk *et al*, *NEJM,* 2001.
- Prostate cancer: prediction of patient outcome after prostatectomy,
  - Singh D*. et al, Cancer Cell,* 2002.
- Malignant gliomas survival: gene expression vs. histological classification,
  - C. Nutt *et al*, *Cancer Res.*, 2003.
- Clinical outcome of breast cancer,
  - L. van't Veer *et al, Nature*, 2002.
- Recurrence of hepatocellaur carcinoma after curative resection,
  - N. Iizuka *et al*, *Lancet*, 2003.
- Tumor vs. normal colon tissues,
  - A. Alon *et al*, *PNAS,* 1999.
- Acute Myeloid vs. Lymphoblastic Leukemia,
  - T. Golub *et al*, *Science,* 1999.

Feature selection techniques

- Standard methods need long and memory intensive computations.
  - – PCA, SVD, ICA,…
- Statistical techniques are much faster, but can produce low accuracy results.
  - – FDA, LDA,…
- Need for hybrid techniques that can take advantage of both approaches.

ILDC-ReGEC

- Simultaneous incremental learning and decremented characterization permit to acquire knowledge on gene grouping during the classification process.

- This technique relies on standard statistical indexes (mean μ and standard deviation σ):

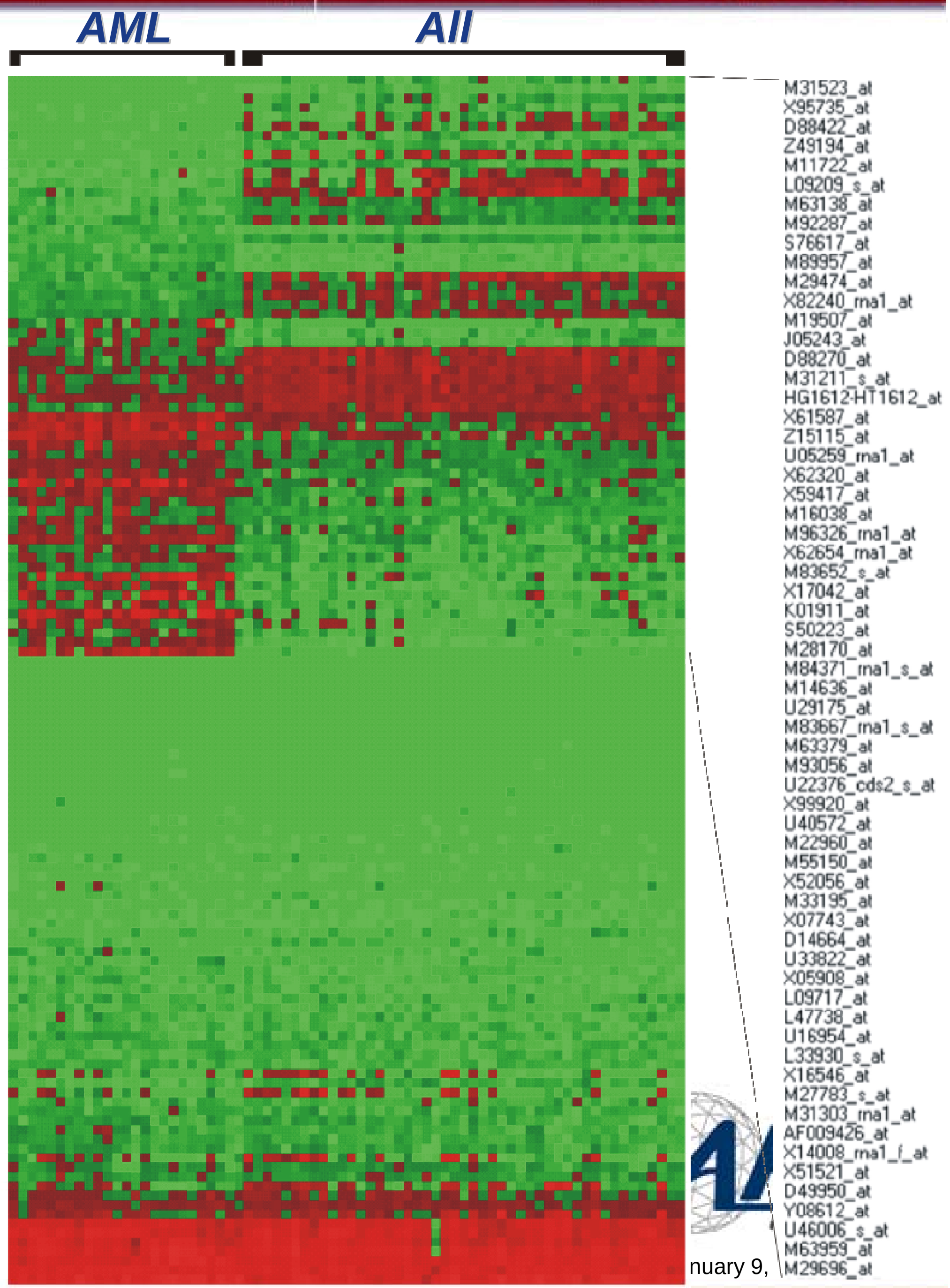
$$F(x_j) = |\frac{\mu_j^+ - \mu_j^-}{\sigma_j^+ + \sigma_j^-}|$$

ILDC-ReGEC: Golub dataset

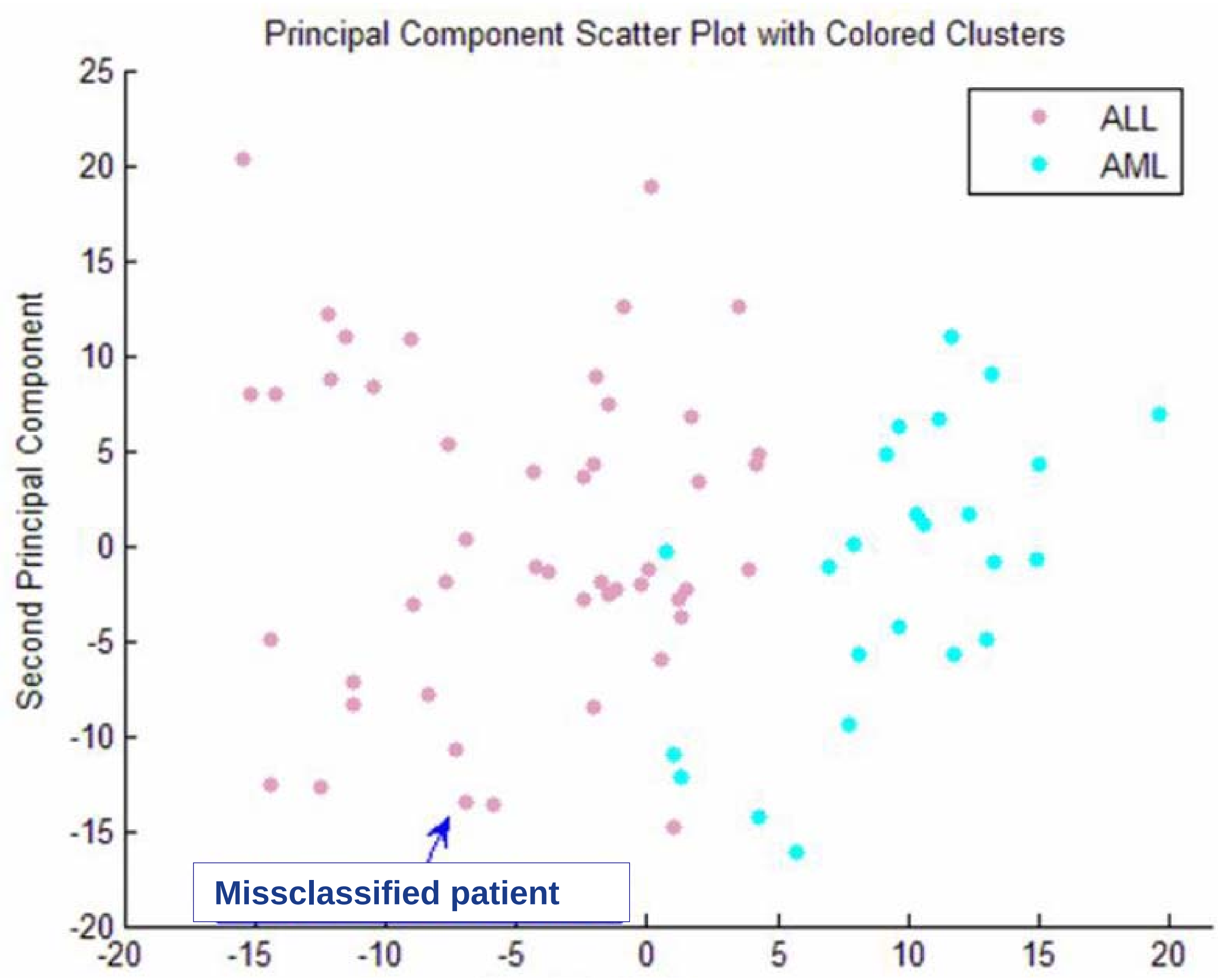
- About 100 genes out of 7129 responsible of discrimination
  - Acute Myeloid Leukemia, and
  - Acute Lymphoblastic Leukemia.

- Selected genes in agreement with previous studies.

- Less then 10 patients, out of 72, needed for training.
  - Classification accuracy: 96.86%



ILDC-ReGEC: Golub dataset



Missclassified patient

- Different techniques agree on the miss-classified patient!

Gene expression analysis

- ILDC-ReGEC
  - Incremental classification with feature selection for microarray datasets.
- Few experiments and genes selected as important for discrimination.

Dataset	*chunk*	*% of train*	*features*	*% of features*
H-BRCA1 *22 x 3226*	6.11	30.55	49.85	1.55
H-BRCA2 *22 x 3226*	4.28	21.40	56.48	1.75
H-Sporadic *22 x 3226*	6.80	34.00	57.15	1.77
Singh *136 x 12600*	6.87	5.63	288.23	2.29
Nutt *50 x 12625*	8.29	18.42	211.66	1.68
Vantveer *98 x 24188*	8.10	9.31	474.35	1.96
Iizuka *60 x 7129*	20.14	37.30	122.63	1.72
Alon *62 x 2000*	5.43	9.70	32.43	1.62
Golub *72 x 7129*	7.25	11.15	95.39	1.34

ILDC-ReGEC: gene expression analysis

Dataset	LLS SVM	KLS SVM	UPCA FDA	SPCA FDA	LUPCA FDA	LSPCA FDA	KUPCA FDA	KUPCA FDA	ILDC ReGEC
H-BRCA1 *22 x 3226*	75.00	72.62	77.38	75.00	76.19	69.05	66.67	52.38	***80.00***
H-BRCA2 *22 x 3226*	84.52	77.38	72.62	79.76	69.05	72.62	64.29	63.10	***85.00***
H-Sporadic *22 x 3226*	73.81	78.57	69.05	75.00	70.24	***79.76***	69.05	69.05	77.00
Singh *136 x 12600*	***91.20***	90.48	n.a.	n.a.	88.74	84.85	n.a.	n.a.	77.86
Nutt *50 x 12625*	72.22	74.60	n.a.	n.a.	67.46	67.46	n.a.	n.a.	***76.60***
Vantveer *98 x 24188*	66.86	66.86	n.a.	n.a.	65.33	64.57	n.a.	n.a.	***68.00***
Iizuka *60 x 7129*	67.10	61.90	n.a.	n.a.	66.67	61.90	n.a.	n.a.	***69.00***
Alon *62 x 2000*	***91.27***	82.14	90.08	89.68	90.08	84.52	90.87	81.75	83.50
Golub *72 x 7129*	96.83	93.65	93.25	93.25	94.44	90.08	92.06	88.10	***96.86***

Conclusions

- ReGEC is a competitive classification method.
- Incremental learning reduces redundancy in training sets and can help avoiding over-fitting.
- Subset selection algorithm provides a constructive way to reduce complexity in kernel based classification algorithms.
- Initial points selection strategy can help in finding regions where knowledge is missing.
- IReGEC can be a starting point to explore very large problems.