High Performance Computing and Networking Institute
National Research Council, Italy

A constructive approach to incremental learning

Mario Rosario Guarracino

October 12, 2006

Acknowledgements

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

Agenda

- Generalized eigenvalue classification
- 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

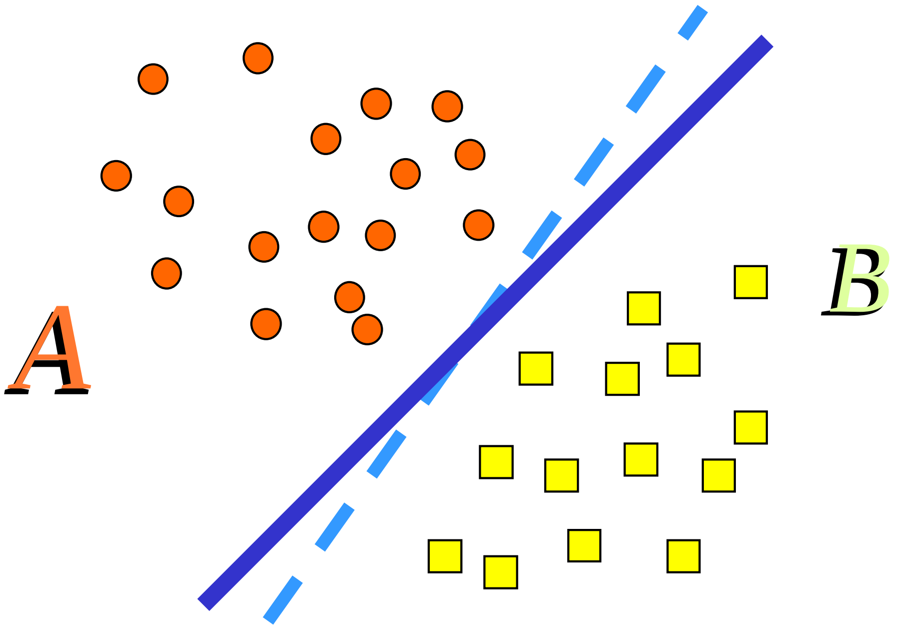
- Many applications in biology and medicine:
  - Tissues that are prone to cancer can be detected with high accuracy.
  - New DNA sequences or proteins can be tracked down to their origins.
  - Identification of new genes or isoforms of gene expressions in large datasets.
  - Analysis and reduction of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

Peculiarity of the problem

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

Linear discriminant planes

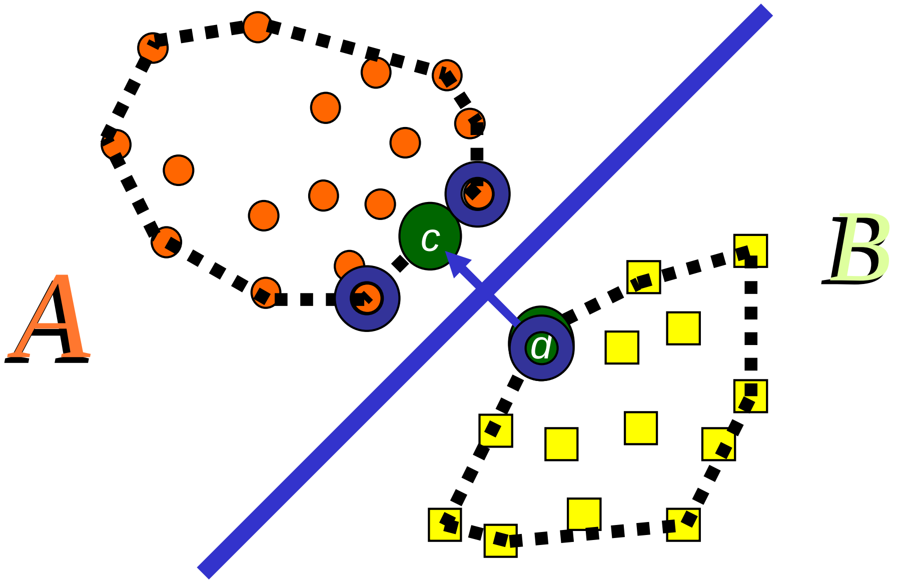
- Consider a binary classification task with points in two linearly separable sets.
  - There exists a plane that classifies all points in the two sets



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

Best plane

- To construct the plane “furthers” from both classes, we examine the *convex hull* of each set.



$$\min_{a} \frac{1}{2} \|c-d\|^2$$

$$c = \sum_{x_i \in A} \alpha_i x_i \quad d = \sum_{x_i \in B} \alpha_i x_i$$

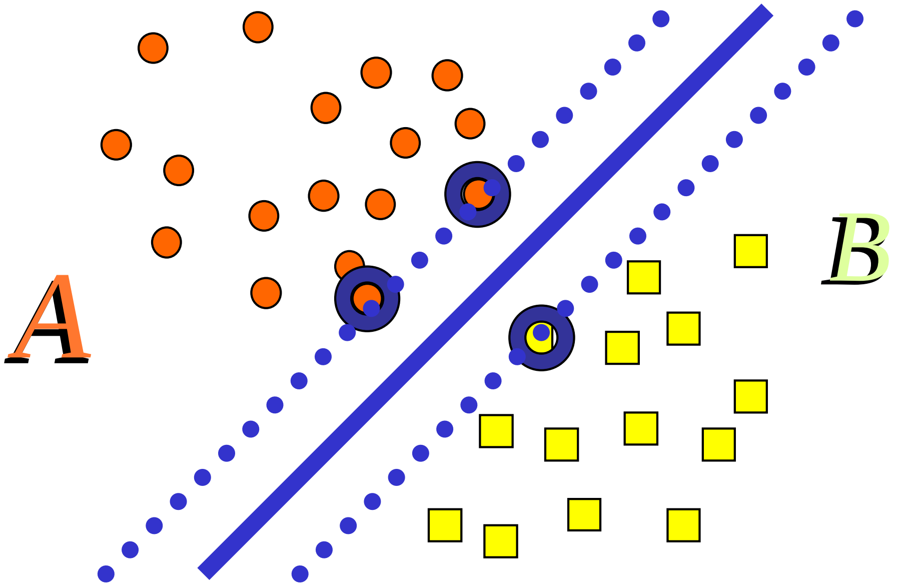
$$s.t. \sum_{x_i \in A} \alpha_i = 1 \quad \sum_{x_i \in B} \alpha_i = 1$$

$$\alpha_i \geq 0$$

- The best plane bisects closest points in the convex hulls.

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.

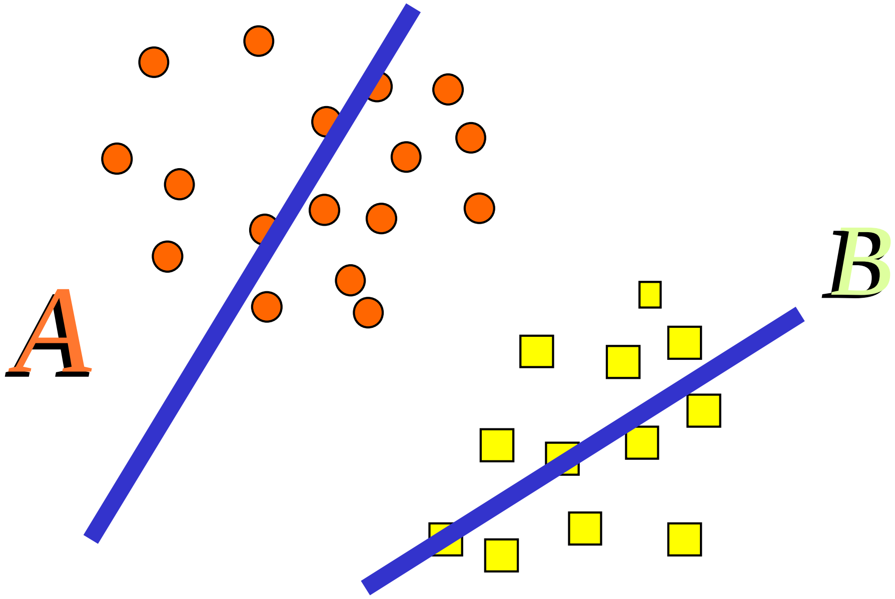
- The training relies on optimization of a quadratic convex cost function, for which many methods are available.
  - Available software includes SVM-Lite and LIBSVM.

- These techniques can be extended to the nonlinear discrimination, embedding the data in a nonlinear space using *kernel functions*.

A different religion

- Mangasarian (2004) showed 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:

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



GEP technique

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

Let:

$$G = [A \quad -e]'[A \quad -e],\; H = [B \quad -e]'[B \quad -e],\; z = [w' \quad \gamma]'$$

Previous equation becomes:

$$\min_{z\in R^m} \frac{z'Gz}{z'Hz}$$

Raleigh quotient of Generalized Eigenvalue Problem

$$Gx=\lambda Hx.$$

GEP technique

Conversely, to find the plane closer to B and further from A we need to solve:

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

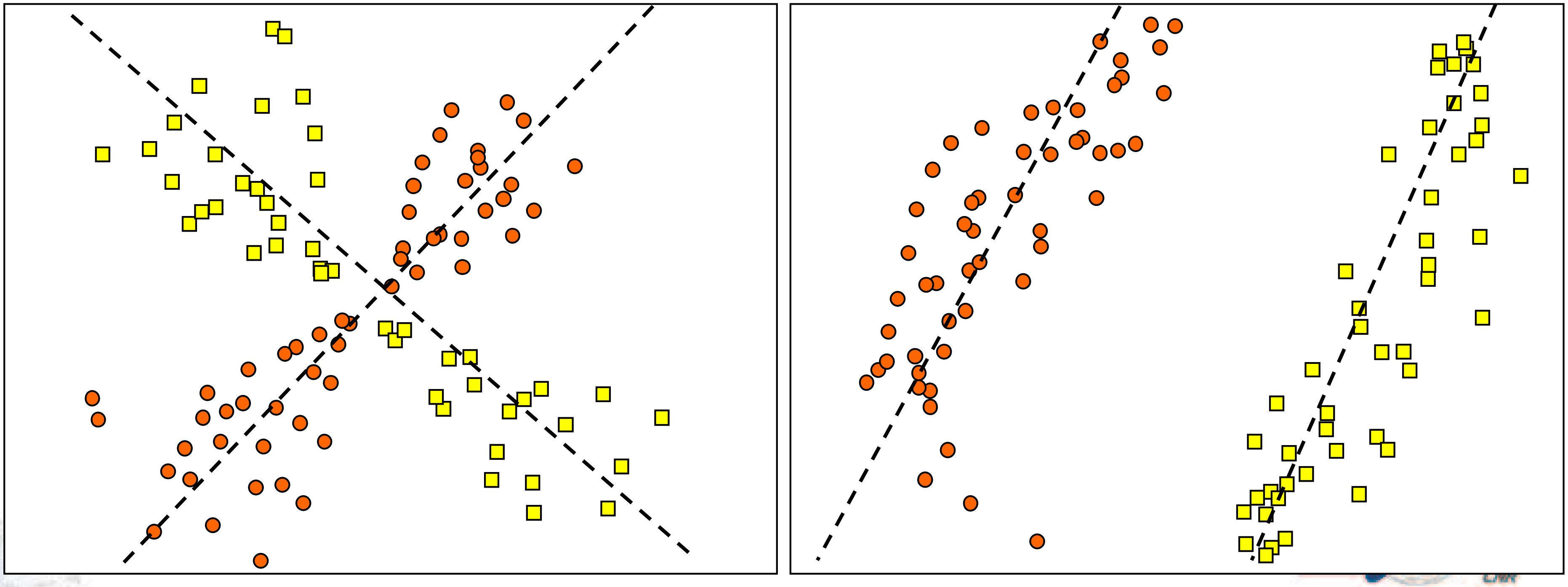
which has the same eigenvectors of the previous problem and reciprocal eigenvalues.

We only need to evaluate the eigenvectors related to min and max eigenvalues of $Gx=\lambda Hx$.

GEP 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 $x'w_1 - \gamma_1 = 0$.



Regularization

- A and B can be rank-deficient.
- G and H are always rank-deficient,
  - the product of matrices of dimension $(n+1 \times n)$ is of rank at least $n \Rightarrow 0/\infty$ eigenvalue.
- Do we need to regularize the problem to obtain a well posed problem?

An useful theorem

Consider GEP $Gx=\lambda Hx$ and the transformed $G_1x=\lambda H_1x$ defined by:

$$G^* = \tau_1 G - \delta_1 H, \quad H^* = \tau_2 H - \delta_2 G,$$

for each choice of scalars τ_1, τ_2, δ_1 and δ_2, such that the 2 × 2 matrix

$$\Omega = \begin{pmatrix} \tau_2 & \delta_1 \\ \delta_2 & \tau_1 \end{pmatrix}$$

is nonsingular.

Then $G^*x=\lambda H^*x$ and $Gx=\lambda Hx$ have the same eigenvectors.

Linear case

- In the linear case, the theorem can be applied. For $\tau_1=\tau_2=1$ and $\delta_1=\delta_2=\delta$, the transformed problem is:

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

- As long as $\delta \neq 1$, matrix Ω is non-degenerate.
- In practice, in each class of the training set, there has to be a number of linearly independent points equal to the number of features.
  - $prob\ (Ker(G) \cap Ker(H) \neq 0) = 0$

Classification accuracy: linear kernel

Dataset	train	dim	ReGEC	GEPSVM	SVM
NDC	300	7	87.60	86.70	89.00
ClevelandHeart	297	13	86.05	81.80	83.60
PimaIndians	768	8	74.91	73.60	75.70
GalaxyBright	2462	14	98.24	98.60	98.30

Accuracy results have been obtained using ten fold cross validation

Nonlinear case

- A standard technique to obtain greater separability between sets is to embed the points into a nonlinear space, via kernel functions, like the *gaussian kernel* :

$$K(x_i, x_j) = e^{-\frac{\|x_i - x_j\|^2}{\sigma}}$$

- Each element of kernel matrix is:

$$K(A, C)_{i,j} = e^{-\frac{\|A_i - C_j\|^2}{\sigma}}$$

where

$$C = \begin{bmatrix} A \\ B \end{bmatrix}$$

Nonlinear case

- Using a gaussian kernel the problem becomes:

$$\min_{w,\gamma\neq 0} \frac{\|K(A,C)u - e\gamma\|^2}{\|K(B,C)u - e\gamma\|^2}$$

- to produce the proximal surfaces:

$$K(x,C)u_1 - \gamma_1 = 0, \quad K(x,C)u_2 - \gamma_2 = 0$$

- The associated GEP involves matrices of the order of the training set and rank at most the number of features.

ReGEC

- Matrices are deeply rank deficient and the problem is ill posed.
- We propose to generate the two proximal surfaces:

$$K(x,C)u_1 - \gamma_1 = 0, \quad K(x,C)u_2 - \gamma_2 = 0$$

solving the problem

$$\min_{w,\gamma\neq 0} \frac{\|K(A,C)u - e\gamma\|^2 + \delta\|\tilde{K}_B u - e\gamma\|^2}{\|K(B,C)u - e\gamma\|^2 + \delta\|\tilde{K}_A u - e\gamma\|^2}$$

where K_A and K_B are main diagonals of $K(A,C)$ and $K(B,C)$.

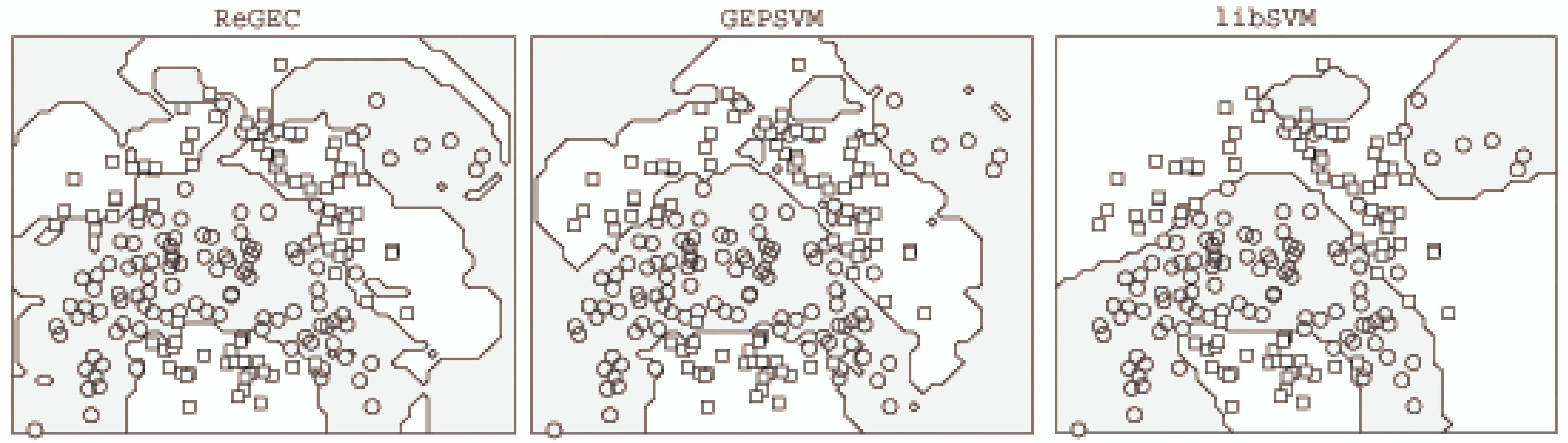
Classification accuracy: gaussian kernel

Dataset	train	test	m	ReGEC	GEPSVM	SVM
Breast-cancer	200	77	9	73.40	71.73	73.49
Diabetis	468	300	8	74.56	74.75	76.21
German	700	300	20	70.26	69.36	75.66
Thyroid	140	75	5	92.76	92.71	95.20
Heart	170	100	13	82.06	81.43	83.05
Waveform	400	4600	21	88.56	87.70	90.21
Flare-solar	666	400	9	58.23	59.63	65.80
Titanic	150	2051	3	75.29	75.77	77.36
Banana	400	4900	2	84.44	85.53	89.15

Accuracy with ten random splits provided by IDA repository

Methods generalization

- The classification surfaces are very tangled.



- Those models are good on original data, but do not generalize well to new data (over-fitting).

How to solve the problem?

Incremental classification

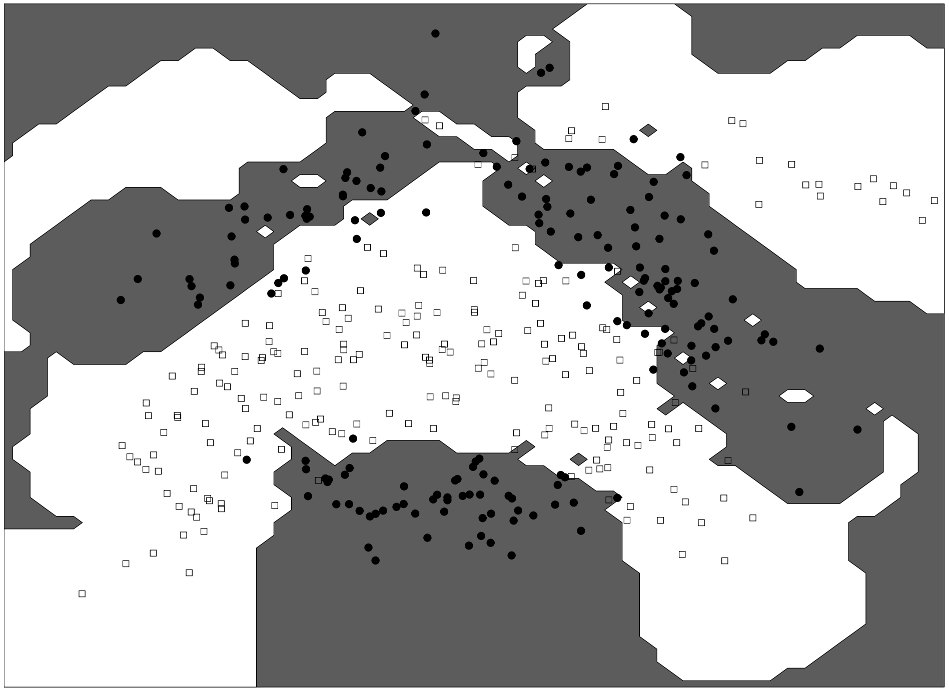
- A possible solution is to find a small and robust subset of the training set that provides comparable accuracy results.
- A smaller set of points reduces the probability of over-fitting the problem.
- A kernel built from a smaller subset is computationally more efficient in predicting new points, compared to kernels that use the entire training set.
- As new points become available, the cost of retraining the algorithm decreases if the influence of the new points is only evaluated by the small subset.

Incremental learning algorithm

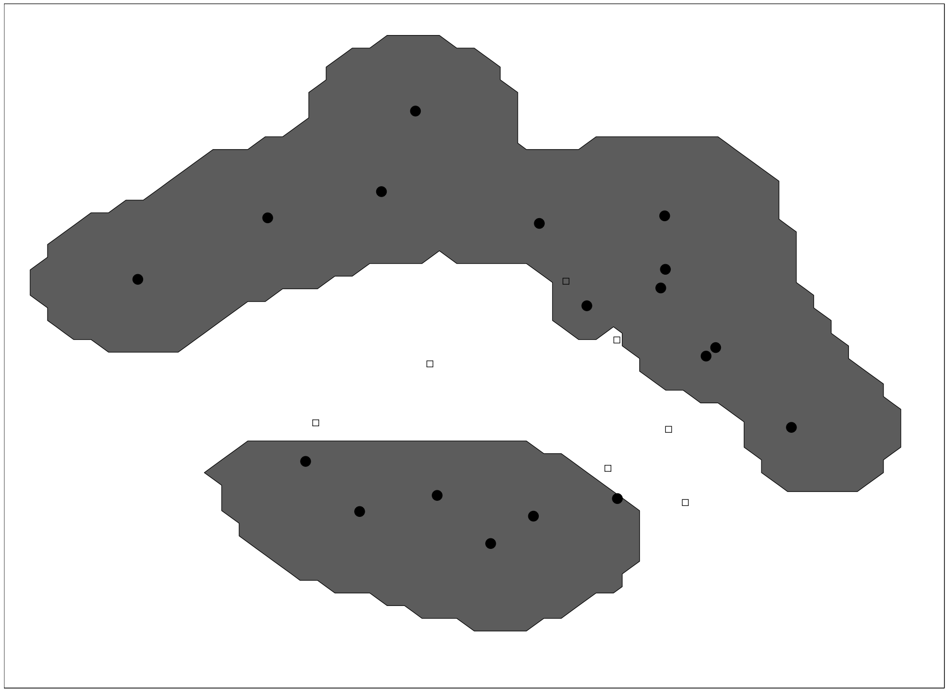
```
$\Gamma_0 = C \setminus C_0$
$\{M_0, Acc_0\} = Classify( C; C_0 )$
$k = 1$
**while** $|\Gamma_k| > 0$ **do**
    $x_k = x : \max_{x \in \{M_k \cap \Gamma_{k-1}\}} \{dist(x, P_{class(x)})\}$
    $\{M_k, Acc_k\} = Classify( C; \{C_{k-1} \cup \{x_k\}\} )$
    **if** $Acc_k > Acc_{k-1}$ **then**
   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

I-ReGEC: Incremental ReGEC

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

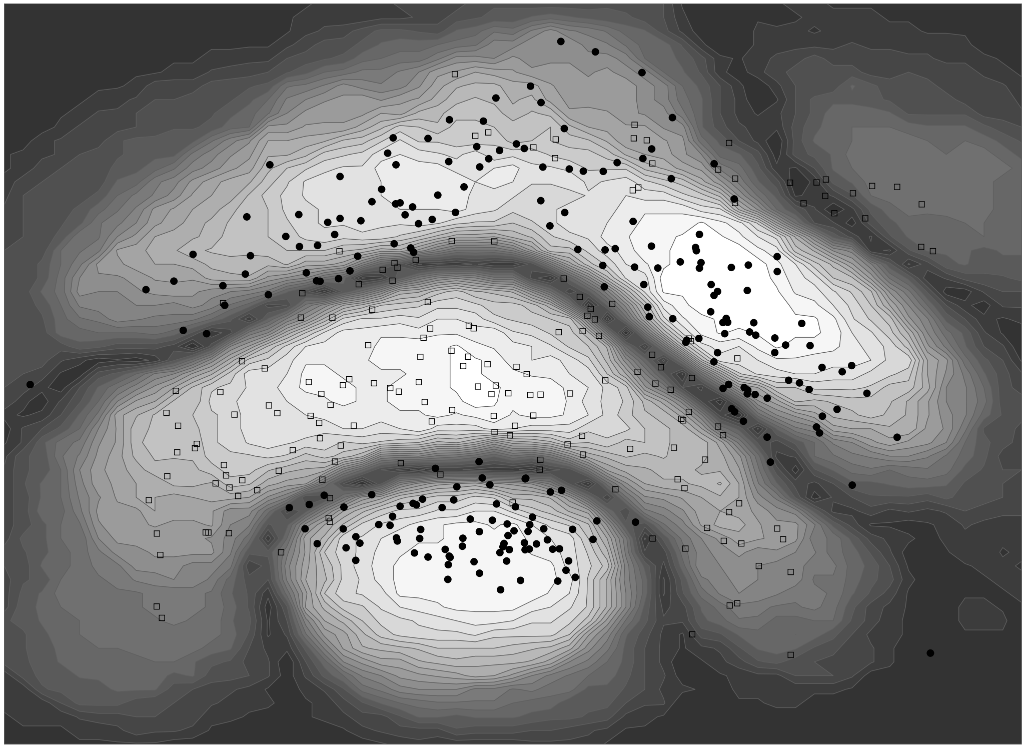
- Unsupervised clustering techniques can be adapted to select initial points.

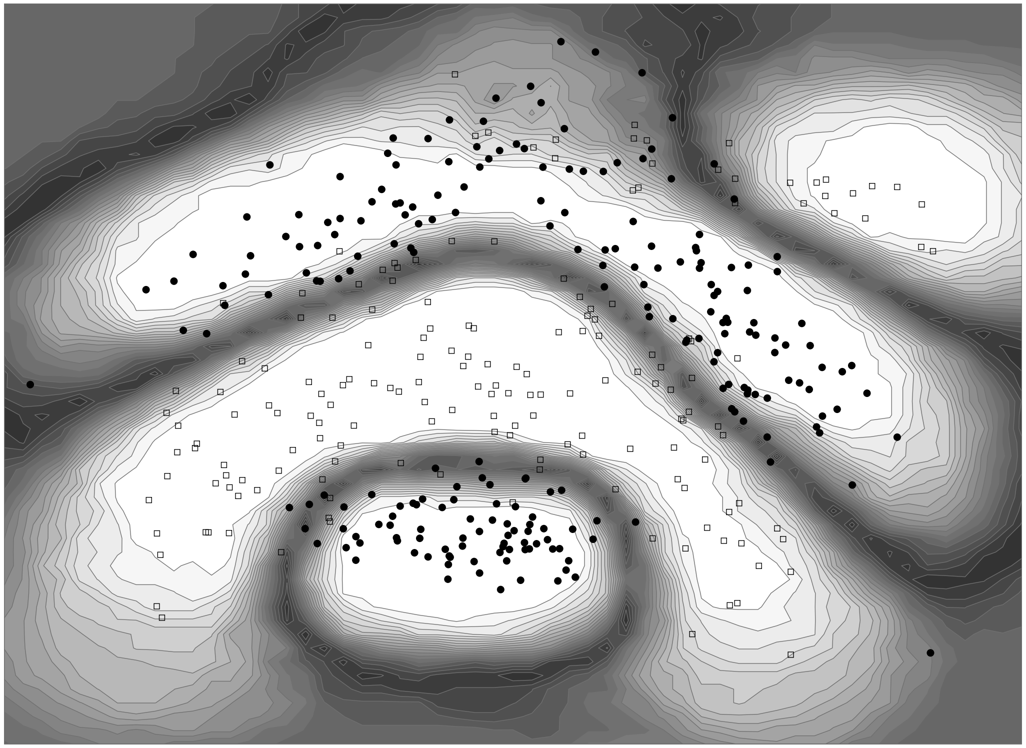
- We compare the classification obtained with k randomly selected starting points for each class, and k points determined by *k-means* method.

- 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

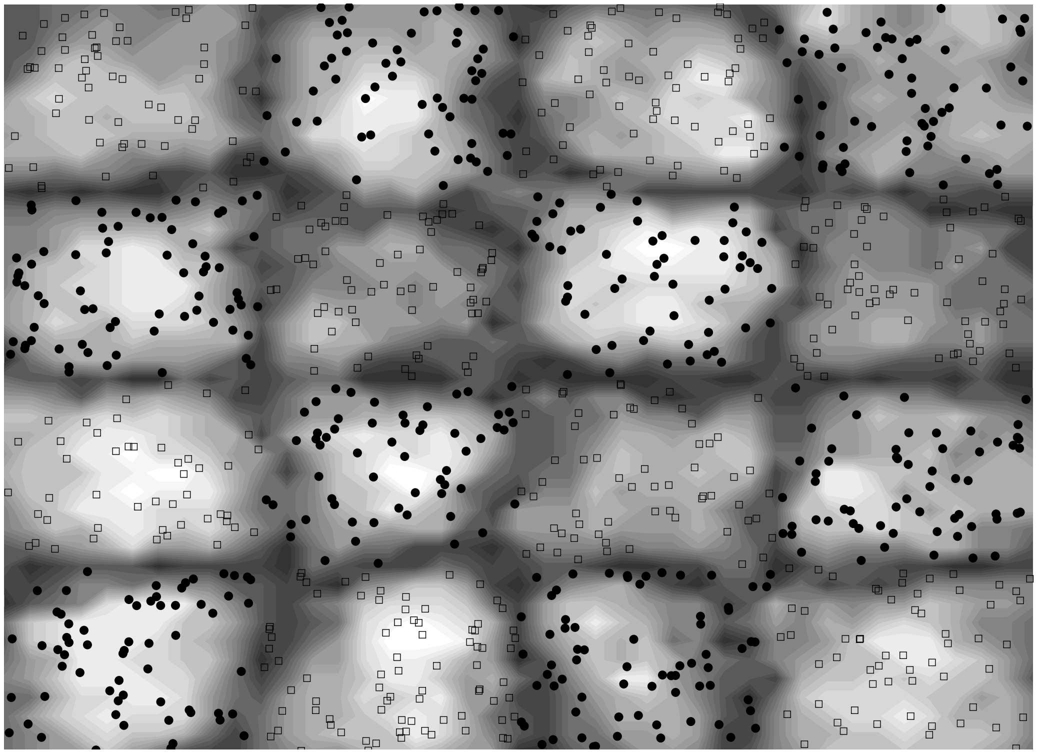
- Starting points C_i chosen:
  - randomly (top),
  - k-means (bottom).
- For each kernel produced by C_i, a set of evenly distributed points x is classified.
  - The procedure is repeated 100 times.
- Let $y_i \in \{1; -1\}$ be the classification based on C_i.
- $y = |\sum y_i|$ estimates the probability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

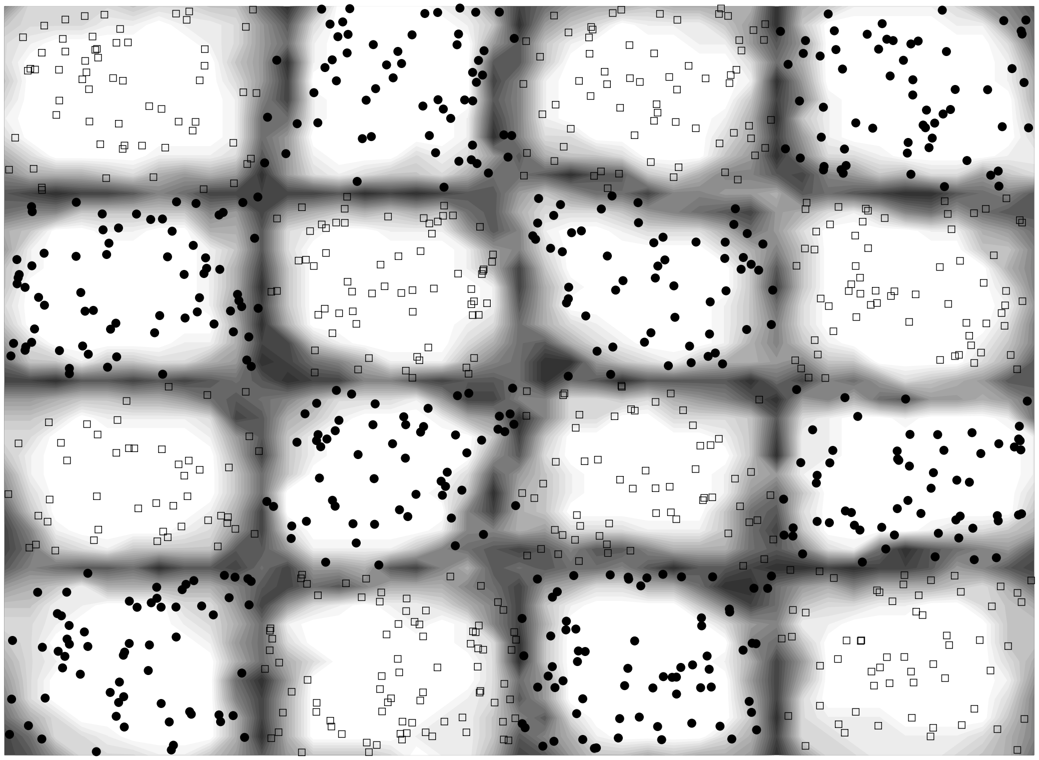




Initial points selection

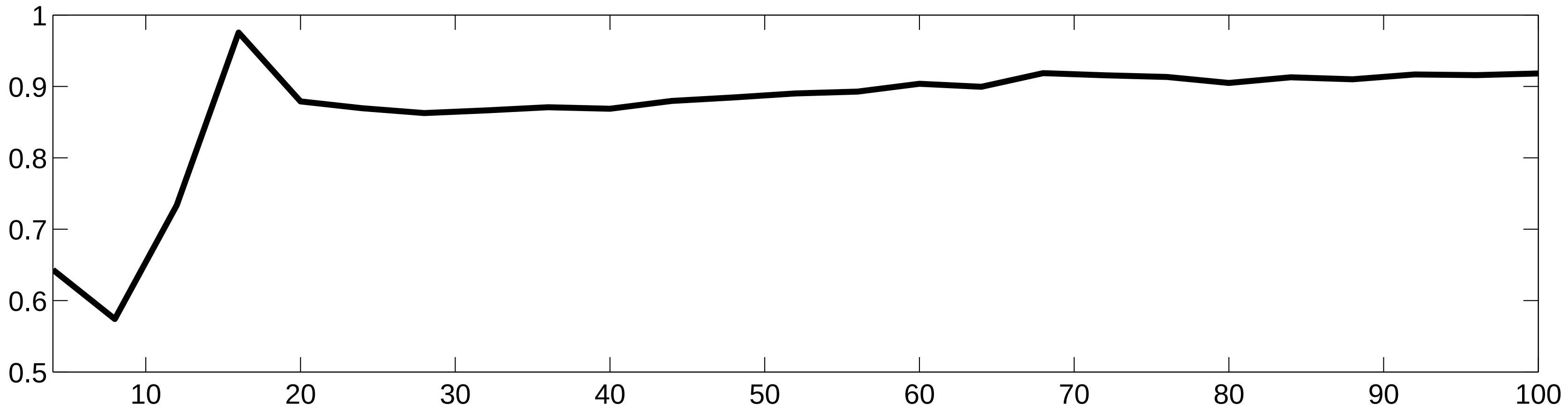
- Starting points C_i chosen:
  - randomly (top),
  - k-means (bottom).
- For each kernel produced by C_i, a set of evenly distributed points x is classified.
  - The procedure is repeated 100 times.
- Let $y_i \in \{1; -1\}$ be the classification based on C_i.
- $y = |\sum y_i|$ estimates the probability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





Initial point selection

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

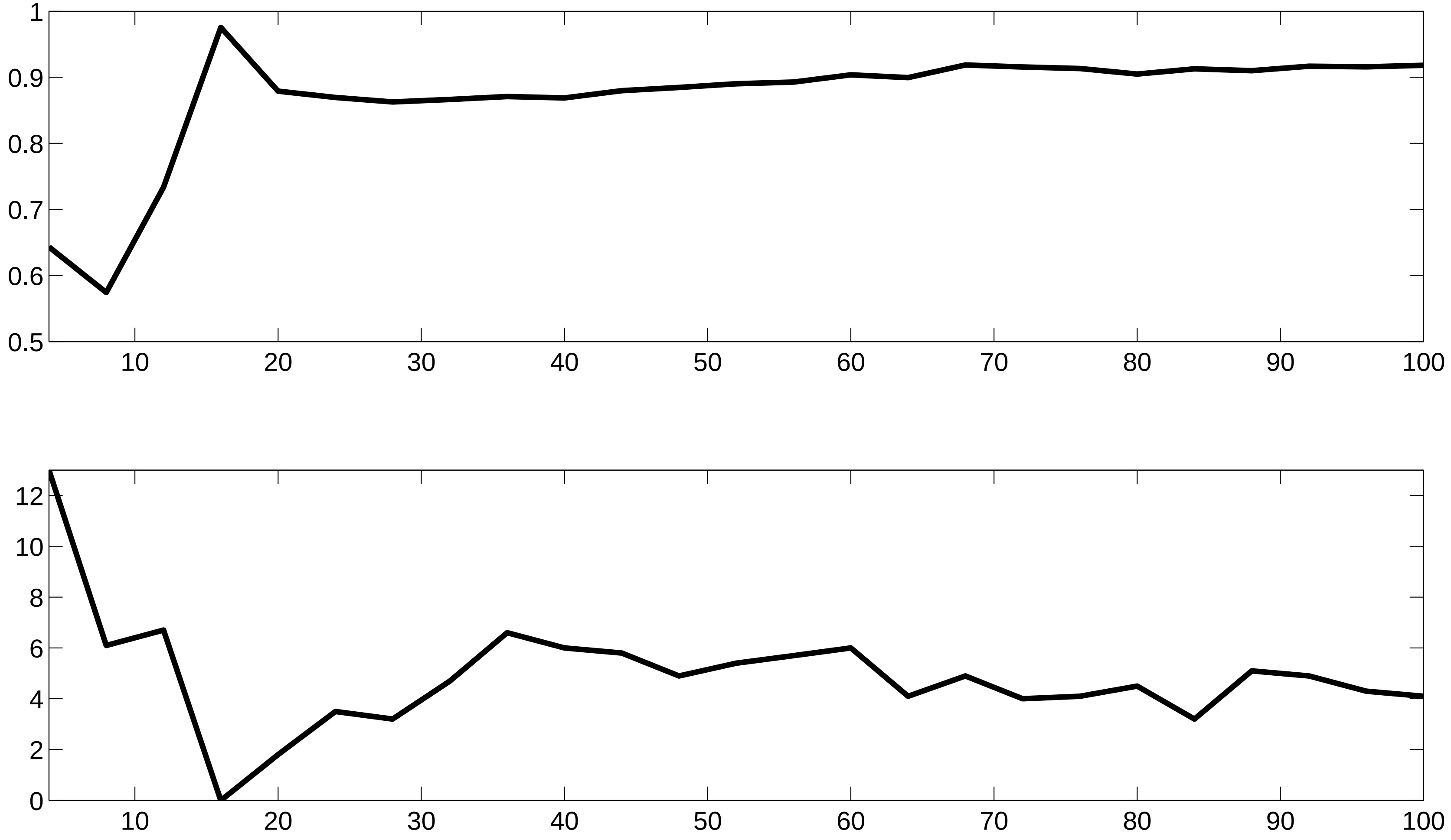


- The graph shows the classification accuracy versus the total number of initial points $2k$ from both classes.

- This result empirically shows that there is a minimum k, with which we reach high accuracy results.

Initial point selection

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



Dataset 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

	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 do not deteriorate selecting fewer training points.
- Classification surfaces can be generalized.

Positive results

- Incremental classification can be applied to different algorithms and still enhances accuracy results

	T.r.a.c.e.	*I-T.r.a.c.e.*
Dataset	*acc (bar)*	*acc (bar)*
Banana	85.06 *(129.35)*	87.26 *(23.56)*
German	69.50 *(268.04)*	72.15 *(34.11)*
Diabetis	67.83 *(185.60)*	72.55 *(9.85)*
Haberman	63.85 *(129.22)*	72.82 *(11.14)*
Bupa	65.80 *(153.80)*	66.21 *(11.79)*
Votes	92.70 *(60.69)*	93.25 *(15.12)*
WPBC	66.00 *(129.35)*	69.78 *(23.56)*
Thyroid	94.77 *(21.57)*	94.55 *(13.41)*
Flare-Solar	60.23 *(68.06)*	65.81 *(4.20)*

courtesy of Claudio Cifarelli

Not so positive results

- There are points in the training set that are not chosen by the method but increase accuracy.
- Block selection does not give any improvement.

Work in progress

- Incremental classification with feature selection for microarray datasets.

Dataset	chunk	% of train	features	% of feature
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

Work in progress

Dataset	***L-LS SVM***	***K-LS SVM***	***U-PCA FDA***	***S-PCA FDA***	***L-U PCA FDA***	***L-S PCA FDA***	***K-U PCA FDA***	***K-U PCA FDA***	***IRegec Golub***
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***

L=linear, K=RBF, U=unsupervised, S=supervised

http://www.esat.kuleuven.be/MACBETH/

Conclusions

- Generalized eigenvalue is a competitive classification method.
- Incremental learning reduces redundancy in training sets and can help to avoid 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.

Questions?

High Performance Computing and Networking Institute
National Research Council, Italy

A constructive approach to incremental learning

Mario.Guarracino@icar.cnr.it
October 12, 2006