High Performance Computing and Networking Institute
National Research Council, Italy

Incremental Classification with Generalized Eigenvalues

Mario Rosario Guarracino

September 17, 2007

Consiglio Nazionale delle Ricerche

People@ICAR

- Researchers
  - **Mario Guarracino**
  - **Pasqua D'Ambra**
  - **Ivan De Falco**
  - **Ernesto Tarantino**
- Associates
  - **Daniela di Serafino (SUN)**
  - **Francesca Perla** (UniParth)
  - **Gerardo Toraldo** (UniNa)
- Fellows
  - **Davide Feminiano**
  - **Salvatore Cuciniello**
- Collaborators
  - **Franco Giannessi (UniPi)**
  - **Claudio Cifarelli (HP)**
  - **Panos Pardalos, Onur Seref (UFL)**
  - **Oleg Prokopyev (U. Pittsburg)**
  - **Giuseppe Trautteur (UniNa)**
  - **Francesca Del Vecchio Blanco (SUN)**
  - **Antonio Della Cioppa (UniSa)**
- Students
  - *Danilo Abbate,*
  - *Francesco Antropoli,*
  - *Giovanni Attratto,*
  - *Tony De Vivo,*
  - *Alessandra Vocca,*

Agenda

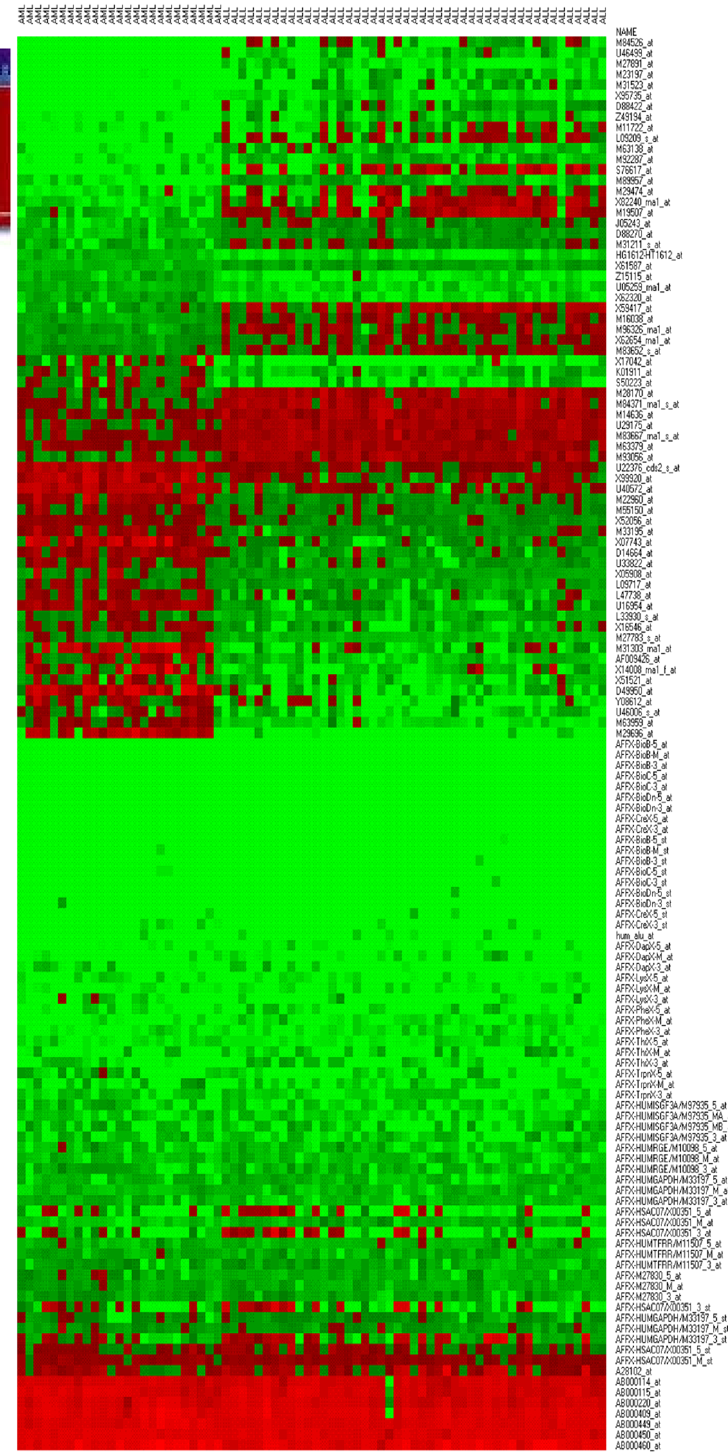
- Generalized eigenvalues classification
- Purpose of incremental learning
- Subset selection algorithm
- Initial points selection
- Accuracy results
- More examples
- Conclusion and future work

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.

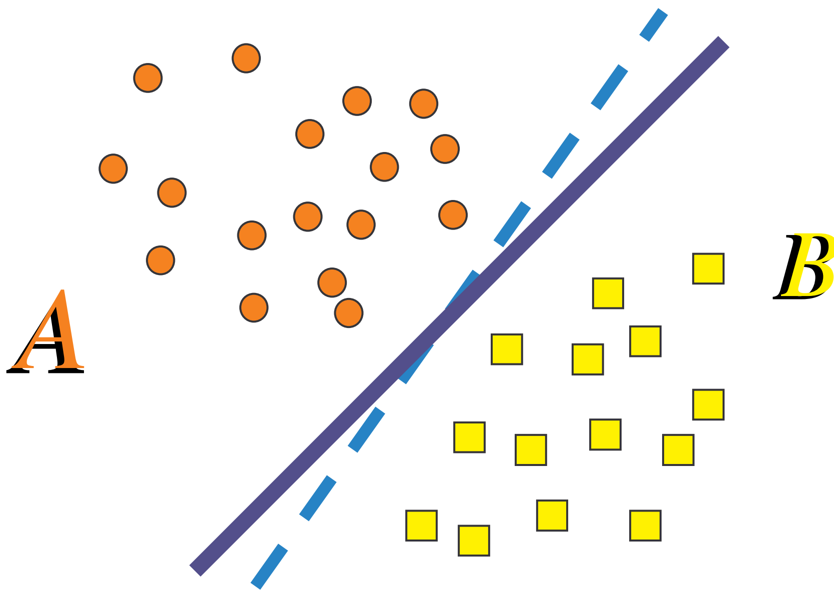
Applications

- 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

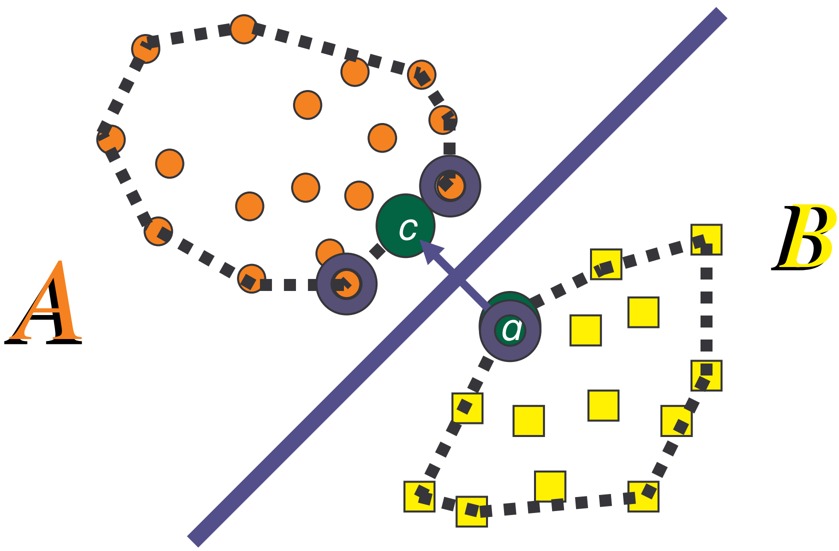
- Consider a binary classification task with points in two linearly separable sets.
  - There exists a plane that classifies all points in the two sets



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

Support vector machines formulation

- To construct the furthest plane from both sets, we examine the *convex hull* of each set.



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

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

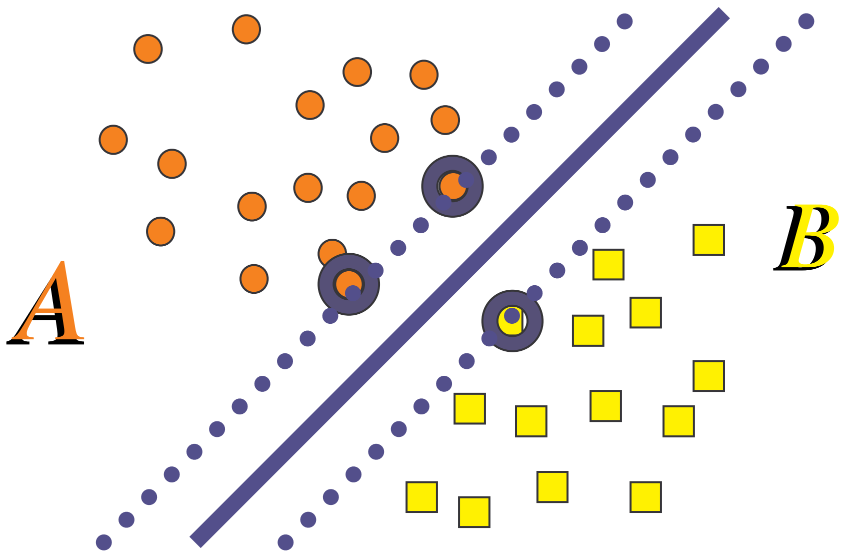
$$s.t. \sum_{x_i \in A} \alpha_i = 1 \quad \sum_{x_i \in B} \beta_i = 1$$

$$\alpha_i, \beta_i \geq 0$$

- The best plane bisects closest points (*support vectors*) in the convex hulls.

Support vector machines dual formulation

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



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

$$s.t.$$

$$Aw + b \geq e$$

$$Bw + b < -e$$

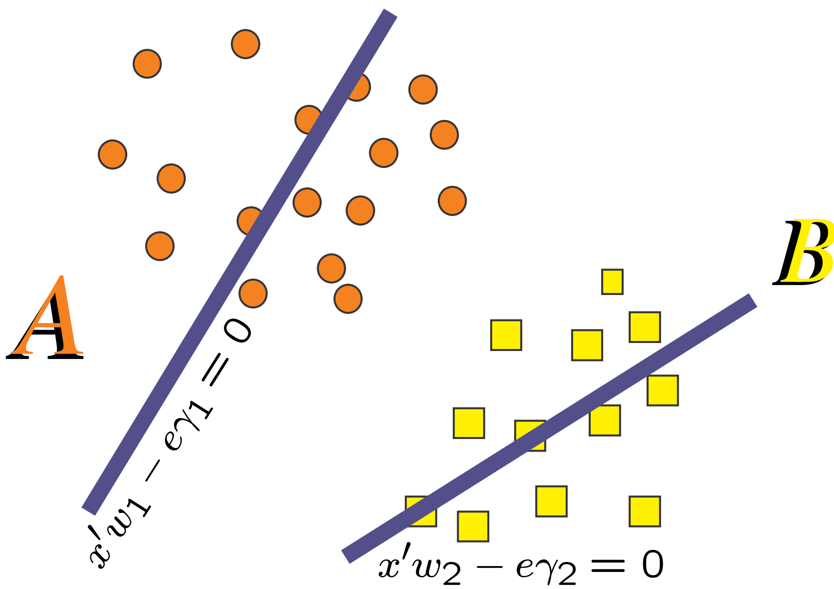
- Support planes are pushed apart until they "bump" into a small set of data points (*support vectors*).

Support Vector Machine features

- 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 do not scale well with the size of the training set.
  - Training 50,000 examples amounts to a Hessian matrix with 2.5 billion elements ~ 20 GB RAM.

A different approach

- The problem can be restated as: find two hyperplanes, each the closest to one set and the furthest from the other.



$$\min_{w_1,\gamma_1 \neq 0} \frac{\|Aw_1 - e\gamma_1\|^2}{\|Bw_1 - e\gamma_1\|^2}$$

$$\min_{w_2,\gamma_2 \neq 0} \frac{\|Bw_2 - e\gamma_2\|^2}{\|Aw_2 - e\gamma_2\|^2}$$

- The binary classification problem can be solved as a generalized eigenvalue computation (GEC).

O. L. Mangasarian and E. W. Wild Multisurface Proximal Support Vector Classification via Generalized Eigenvalues. Data Mining Institute Tech. Rep. 04-03, June 2004.

GEC method

$$\min_{w,\gamma\neq 0} \frac{\|Aw - e\gamma\|^2}{\|Bw - e\gamma\|^2} = \min_{w,\gamma\neq 0} \frac{\|[A \quad -e][\omega' \quad \gamma']'\|^2}{\|[B \quad -e][\omega' \quad \gamma']'\|^2}.$$

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:

$$Gx = \lambda Hx.$$

GEC method

Conversely, the plane closer to B and furthest from A:

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

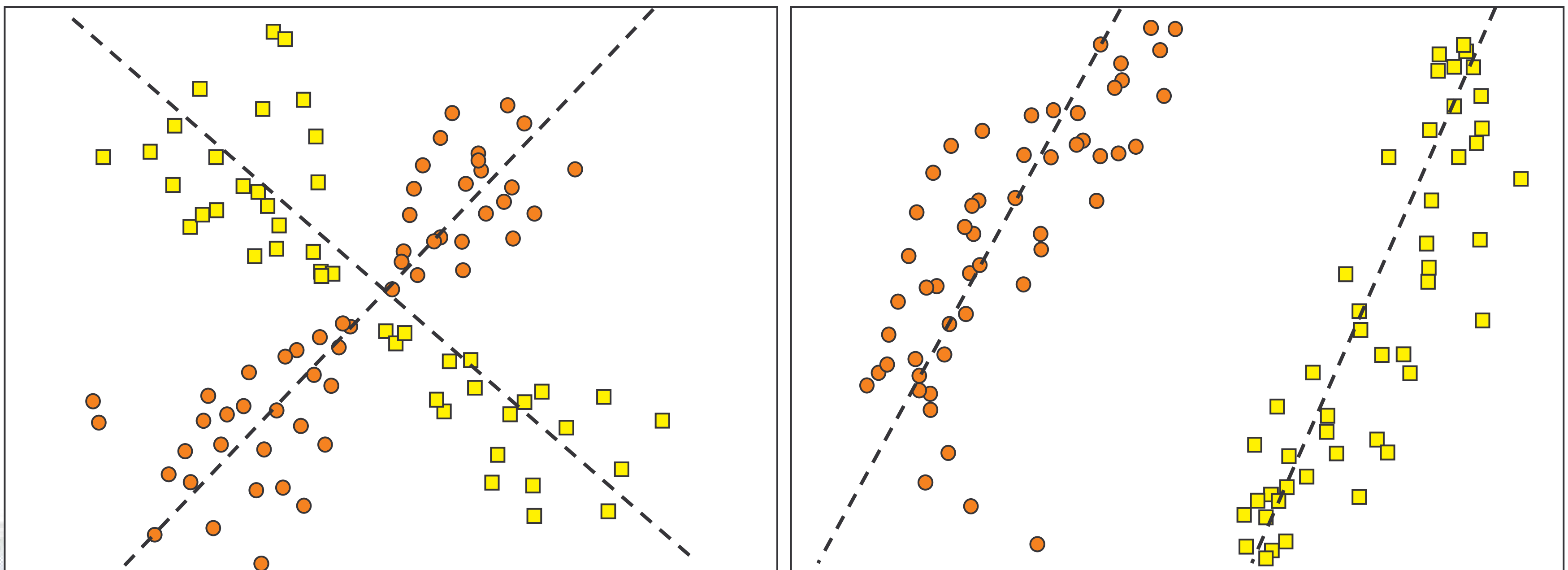
- Same eigenvectors of the previous problem and reciprocal eigenvalues.
- We only need to evaluate the eigenvectors related to minimum and maximum eigenvalues of $Gx=\lambda Hx$.

ICAR CNR

GEC method

Let $[w_1\ \gamma_1]$ and $[w_2\ \gamma_2]$ be eigenvectors associated to min and max eigenvalues of $Gx = \lambda Hx$:

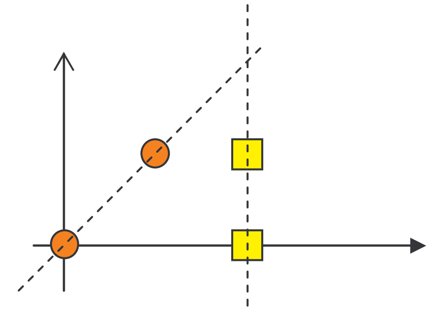
- $a \in A$ closer to $x'w_1 - \gamma_1 = 0$ than to $x'w_2 - \gamma_2 = 0$,
- $b \in B$ closer to $x'w_2 - \gamma_2 = 0$ than to $x'w_1 - \gamma_1 = 0$.



Example

Let:

$$A = \begin{bmatrix} 2 & 0 \\ 2 & 1 \end{bmatrix}, \quad B = \begin{bmatrix} 0 & 0 \\ 1 & 1 \end{bmatrix},$$



Set $G=[A\ -e]'\ [A\ -e]$ and $H=[B\ -e]'\ [B\ -e]$, we obtain:

$$G = \begin{bmatrix} 8 & 2 & -4 \\ 2 & 1 & -1 \\ -4 & -1 & 2 \end{bmatrix}, \quad H = \begin{bmatrix} 1 & 1 & -1 \\ 1 & 1 & -1 \\ -1 & -1 & 2 \end{bmatrix}.$$

Minimum and maximum eigenvalues of $Gx = \lambda Hx$ are $\lambda_1 = 0$ and $\lambda_3 = \infty$ and the corresponding eigenvectors:

$$x_1=[1\ 0\ 2], \quad x_3=[1\ -1\ 0].$$

The resulting planes are $x - 2 = 0$ and $x - y = 0$.

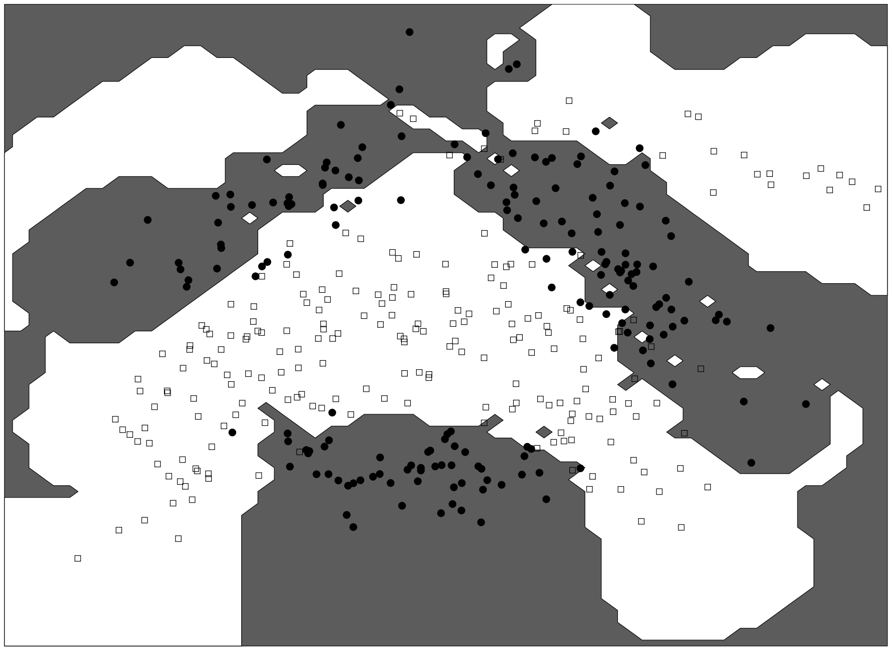
Classification accuracy: linear kernel

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 using ten fold cross validation

Nonlinear case

- When sets are not linearly separable, nonlinear discrimination is needed.



- Data is nonlinearly transformed in another space to increase separability, and linear discrimination is found in that space.

Nonlinear case

- A standard technique is to transform 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 the *kernel matrix* is:

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

where

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

K. Bennett and O. Mangasarian, *Robust Linear Programming Discrimination of Two Linearly Inseparable Sets*, Optimization Methods and Software, 1, 23-34, 1992.

Nonlinear case

- Using the Gaussian kernel the GEC problem can be formulated:

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

in order to evaluate the proximal surfaces:

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

the associated GEC is ill posed.

ReGEC method

- To regularize the problem, generate the two proximal surfaces:

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

solving:

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

where $\tilde{K}_A$ and $\tilde{K}_B$ are main diagonals of $K(A,C)$ and $K(B,C)$.

M. R. Guarracino, C. Cifarelli, O. Seref, P. M. Pardalos, *A Classification Method based on Generalized Eigenvalue Problems*, Optimization Methods and Software, 2007.

ReGEC algorithm

% Let A $\in R^{m\times s}$ and B $\in R^{n\times s}$
% be the training points in each class.
% Choose appropriate δ and $\sigma \in R$

$C = [A;B];$

% Build G and H matrices

$g = [K(A, C, \sigma), -ones(m, 1)];$

$h = [K(B, C, \sigma), -ones(n, 1)];$

$G = g' \times g;$

$H = h' \times h;$

% Regularize the problem

$G^* = G + \delta \times diag(H);$

$H^* = H + \delta \times diag(G);$

% Compute the hyperplanes V(:,1) and V(:,2)

$[V,D] = eig(G^*;H^*);$

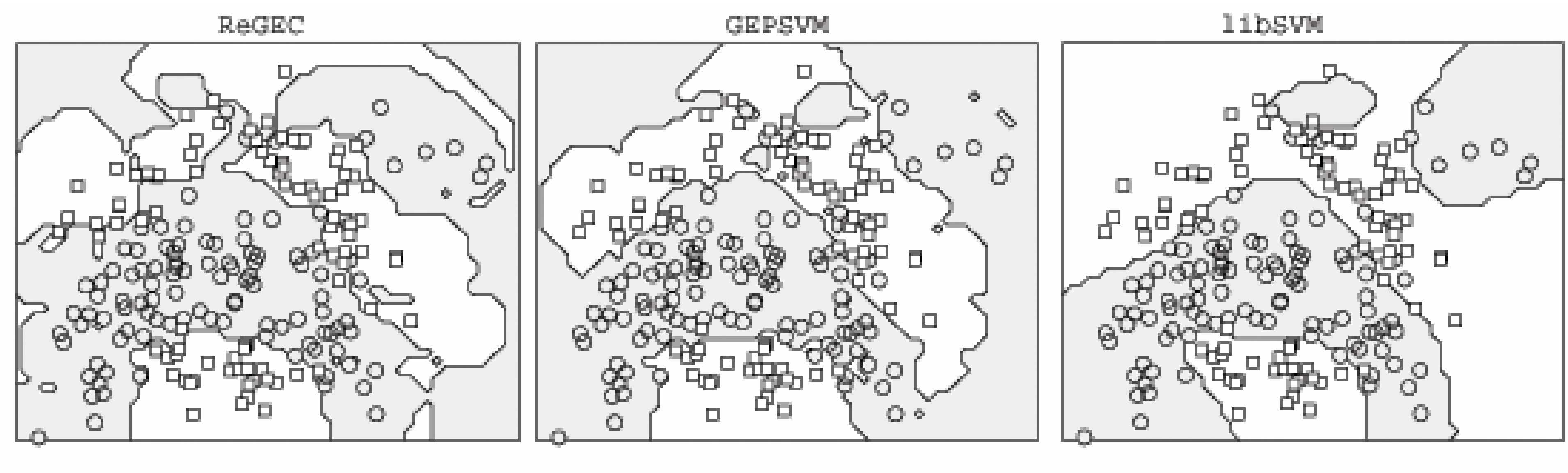
Classification accuracy: gaussian kernel

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

Generalizability of the methods

- The classification surfaces can be 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.

I-ReGEC: Incremental learning

$\Gamma_0 = C \setminus C_0$

$\{M_0, Acc_0\} = Classify(\ C;\ C_0\ )$

$k = 1$

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 learning

$\Gamma_0 = C \setminus C_0$

$\{M_0, Acc_0\} = Classify(\, C;\, C_0\,)$

$k = 1$

while $|\Gamma_k| > 0$ **do**

$\quad x_k = x : \max_{x \in \{M_k \cap \Gamma_{k-1}\}} \{dist(x, P_{class(x)})\}$

$\quad \{M_k, Acc_k\} = Classify(\, C;\, \{C_{k-1} \cup \{x_k\}\}\,)$

$\quad$ **if** $Acc_k > Acc_{k-1}$ **then**

$\qquad C_k = C_{k-1} \cup \{x_k\}$

$\qquad k = k + 1$

$\quad$ **end if**

$\quad \Gamma_k = \Gamma_{k-1} \setminus \{x_k\}$

end while

I-ReGEC: Incremental learning

$\Gamma_0 = C \setminus C_0$

$\{M_0, Acc_0\} = Classify(\, C;\, C_0 \,)$

$k = 1$

while $|\Gamma_k| > 0$ **do**

$\qquad x_k = x : \max_{x \in \{M_k \cap \Gamma_{k-1}\}} \{dist(x, P_{class(x)})\}$

$\qquad \{M_k, Acc_k\} = Classify(\, C;\, \{C_{k-1} \cup \{x_k\}\} \,)$

$\qquad$ **if** $Acc_k > Acc_{k-1}$ **then**

$\qquad\qquad C_k = C_{k-1} \cup \{x_k\}$

$\qquad\qquad k = k + 1$

$\qquad$ **end if**

$\qquad \Gamma_k = \Gamma_{k-1} \setminus \{x_k\}$

end while

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**

        $C_k = C_{k-1} \cup \{x_k\}$

        $k = k + 1$

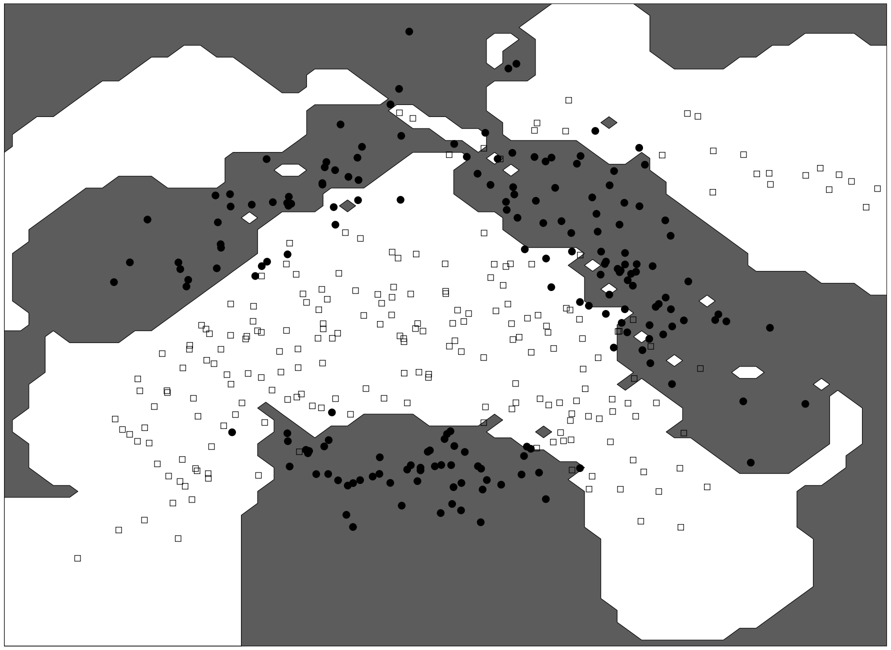
    end if

    $\Gamma_k = \Gamma_{k-1} \setminus \{x_k\}$

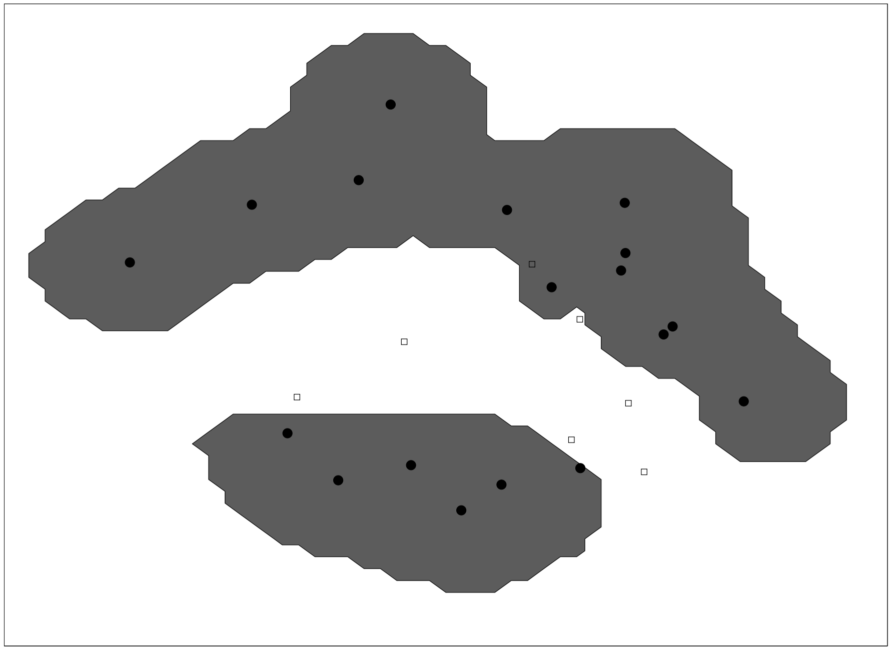
end while

I-ReGEC: Incremental ReGEC

ReGEC accuracy=84.44



I-ReGEC accuracy=85.49



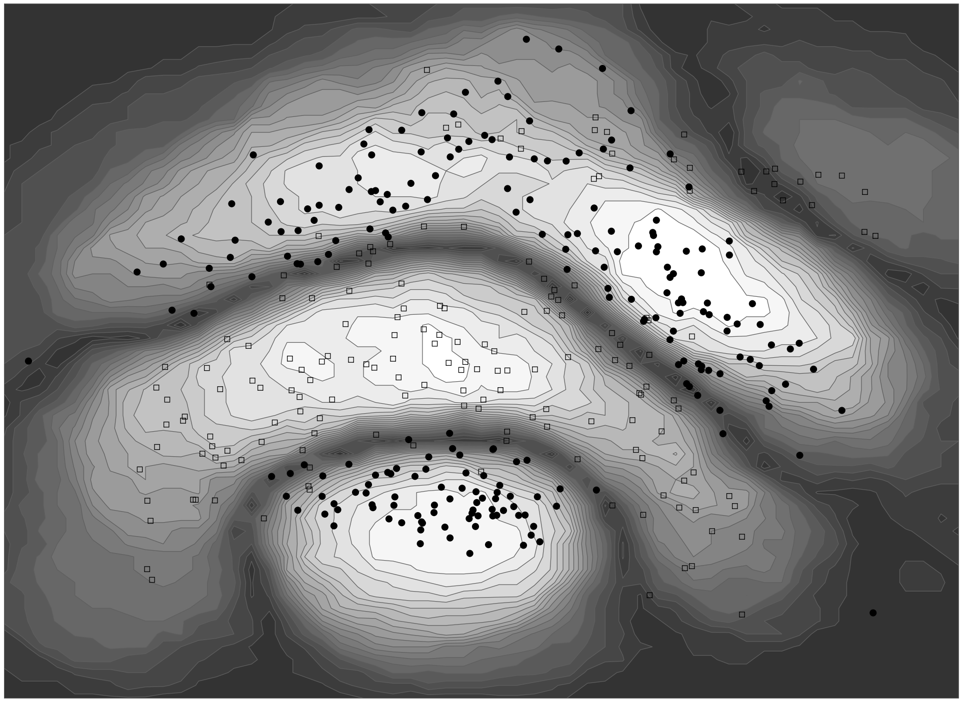
- When ReGEC algorithm is trained on all points, surfaces are affected by noisy points (*left*).
- I-ReGEC achieves clearly defined boundaries, preserving accuracy (*right*).
  - Less then 5% of points needed for training!

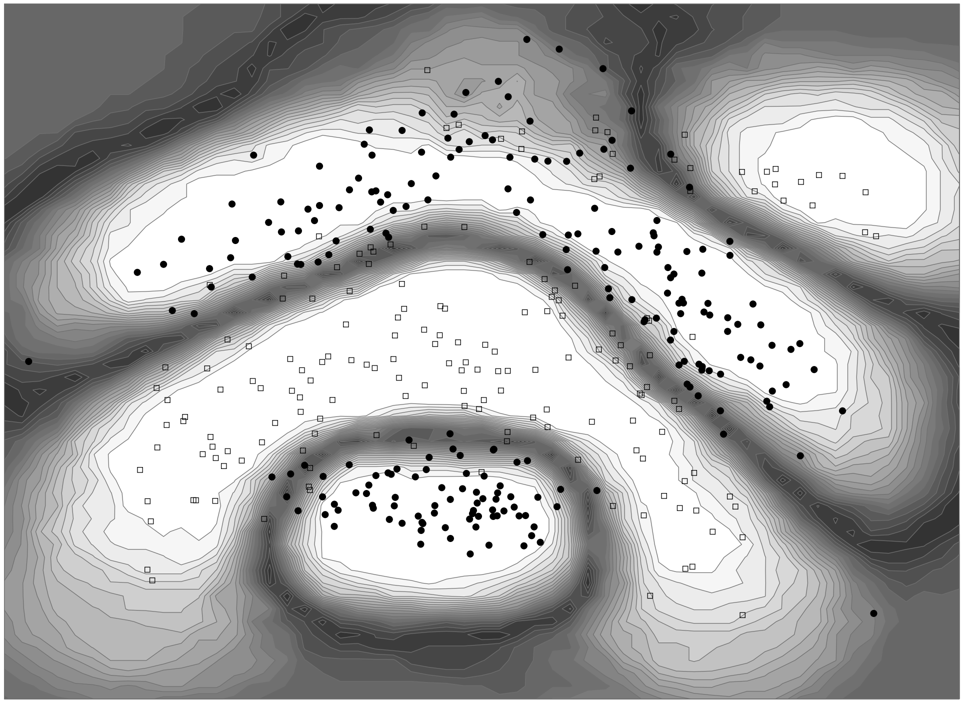
Initial points selection

- 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

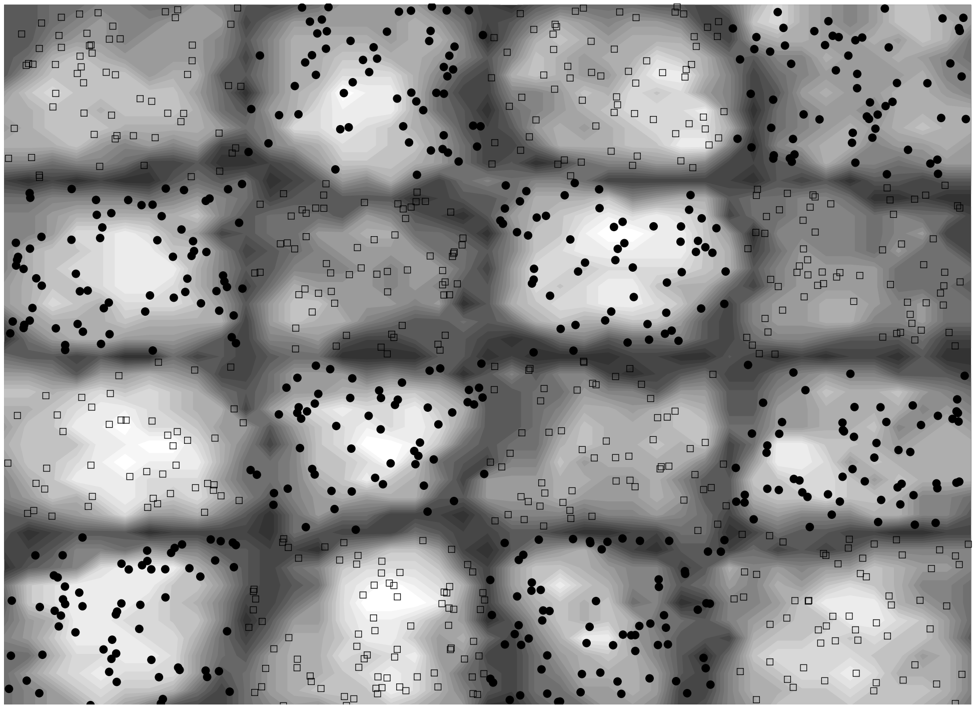
- Starting points C_i chosen:
  - randomly (top),
  - k-means (bottom).
- For each kernel produced by C_i, a set of evenly distributed points x is classified.
  - The procedure is repeated 100 times.
- Let $y_i \in \{1; -1\}$ be the classification based on C_i.
- $y = |\sum y_i|$ estimates the probability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

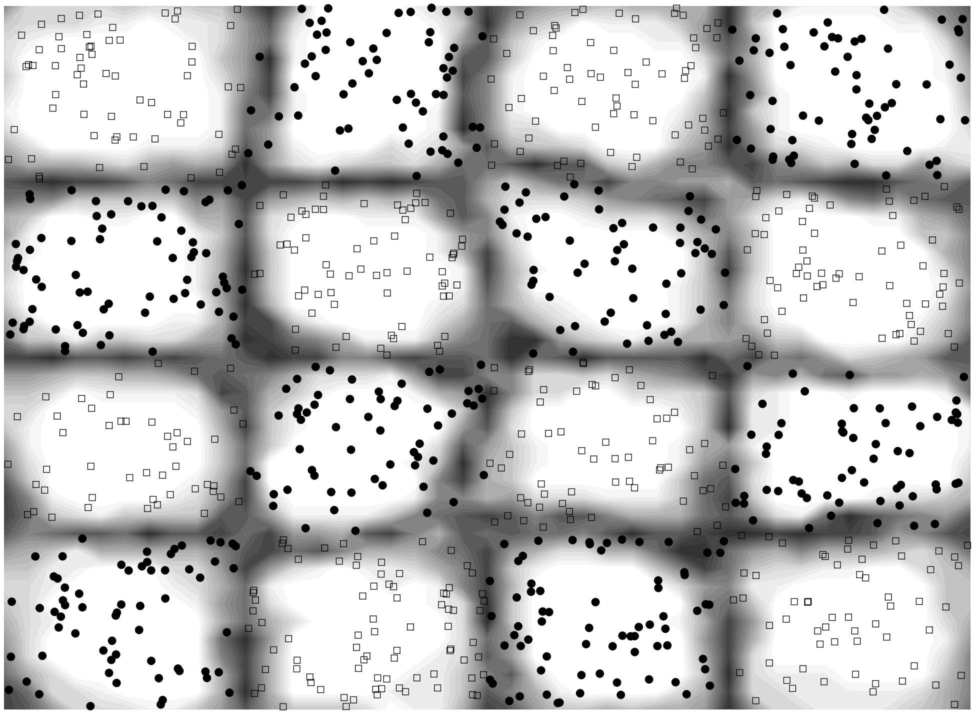




Initial points selection

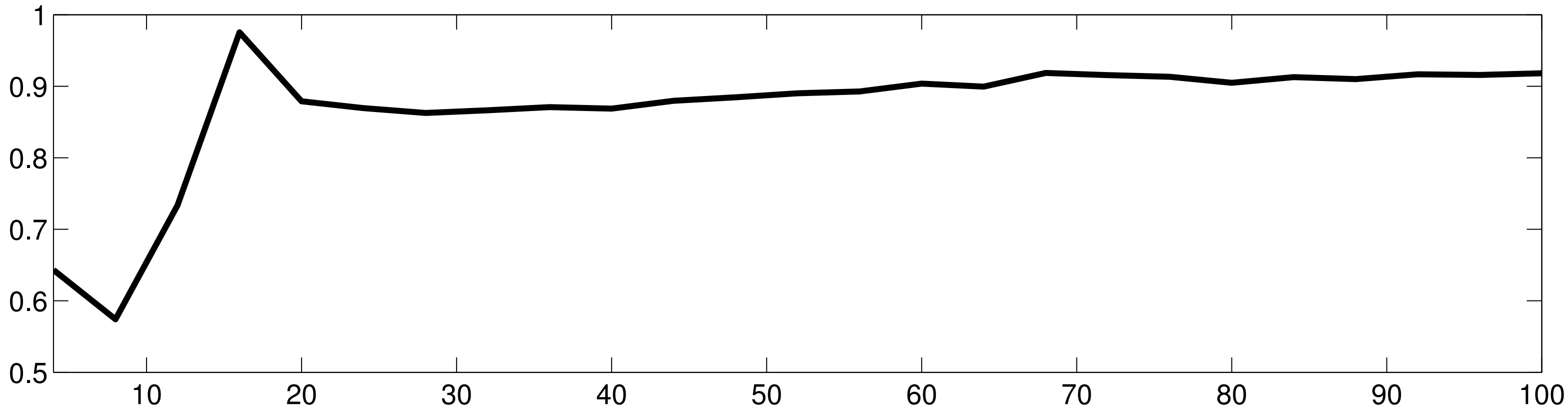
- Starting points C_i chosen:
  - randomly (top),
  - k-means (bottom).
- For each kernel produced by C_i, a set of evenly distributed points x is classified.
  - The procedure is repeated 100 times.
- Let $y_i \in \{1; -1\}$ be the classification based on C_i.
- $y = |\sum y_i|$ estimates the probability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





Initial point selection

- Effect on classification accuracy of increasing initial points with *k-means* on Chessboard dataset (*higher is better*).

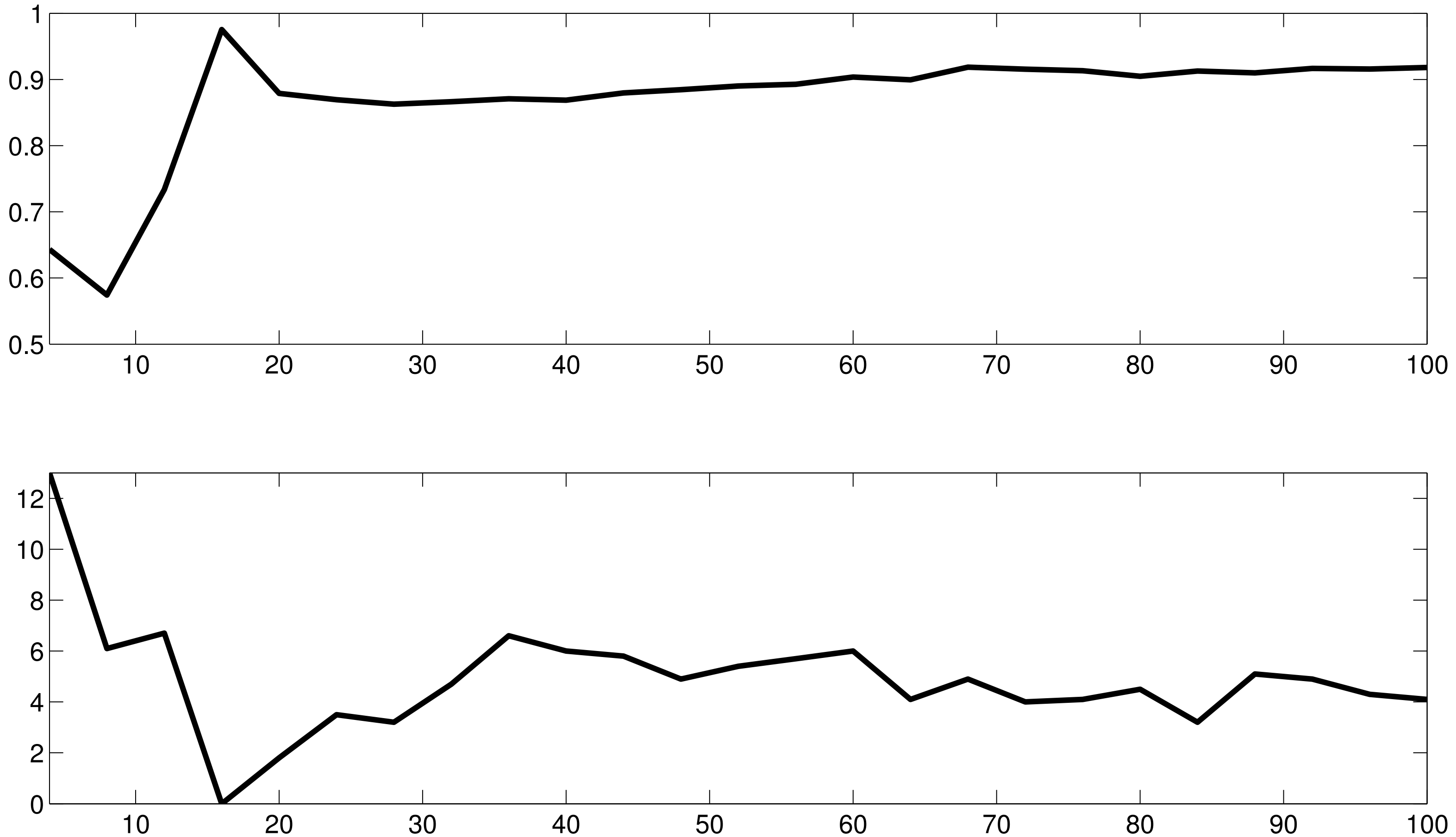


- The graph shows the classification accuracy versus the total number of initial points $2k$ from both classes.

- 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 *(lower is better)*.



Dataset reduction

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

	I-ReGEC		
Dataset	*train*	*chunk*	*% of train*
Banana	400	15.70	3.92
German	700	29.09	4.15
Diabetis	468	16.63	3.55
Haberman	275	7.59	2.76
Bupa	310	15.28	4.92
Votes	391	25.90	6.62
WPBC	99	42.15	4.25
Thyroid	140	12.40	8.85
Flare-solar	666	9.67	1.45

Accuracy results

- Classification accuracy with incremental technique 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 do not deteriorate selecting fewer training points.
- Classification surfaces can be generalized.

Positive results

- Incremental classification can enhance accuracy results of different algorithms.

	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)*

C. Cifarelli, L. Nieddu, O. Seref, P. M. Pardalos. *K-T.R.A.C.E: A kernel k-means procedure for classification*. COR 2007

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. ***(22 patients, 3226 genes)***
- Prostate cancer: prediction of patient outcome after prostatectomy,
  - Singh D. *et al, Cancer Cell,* 2002. ***(136 patients, 12600 genes)***
- Malignant gliomas survival: gene expression vs. histological classification,
  - C. Nutt *et al*, *Cancer Res.*, 2003. (***50 patients, 12625 genes)***
- Clinical outcome of breast cancer,
  - L. van't Veer *et al, Nature*, 2002. ***(98 patients, 24188 genes)***
- Recurrence of hepatocellaur carcinoma after curative resection,
  - N. Iizuka *et al*, *Lancet*, 2003. ***(60 patients, 7129 genes)***
- Tumor vs. normal colon tissues,
  - A. Alon *et al*, *PNAS,* 1999. ***(62 patients, 2000 genes)***
- Acute Myeloid vs. Lymphoblastic Leukemia,
  - T. Golub *et al*, *Science,* 1999. ***(72 patients, 7129 genes)***

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

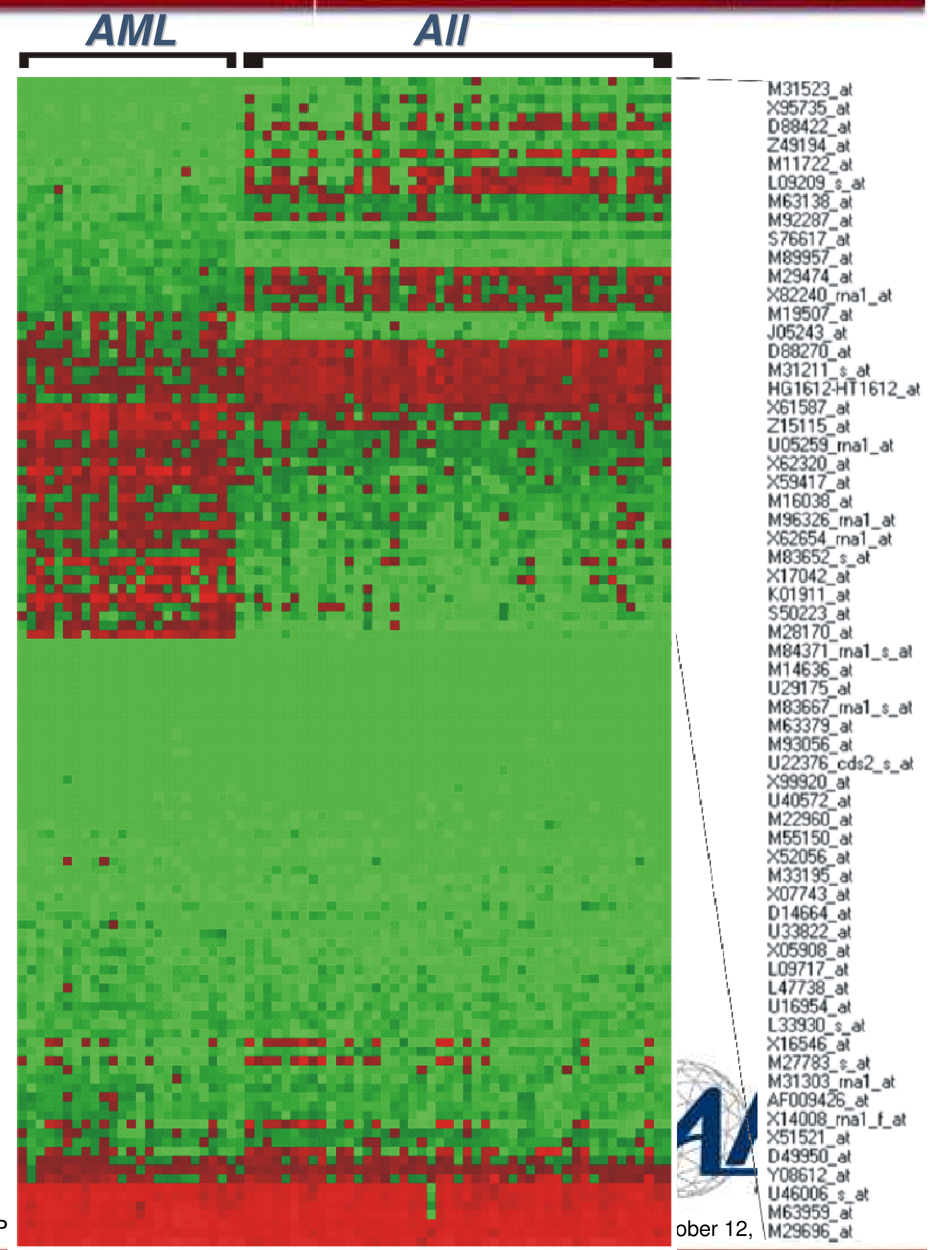
- Simultaneous incremental learning and decremented characterization permit to acquire knowledge on gene grouping during the classification process.

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

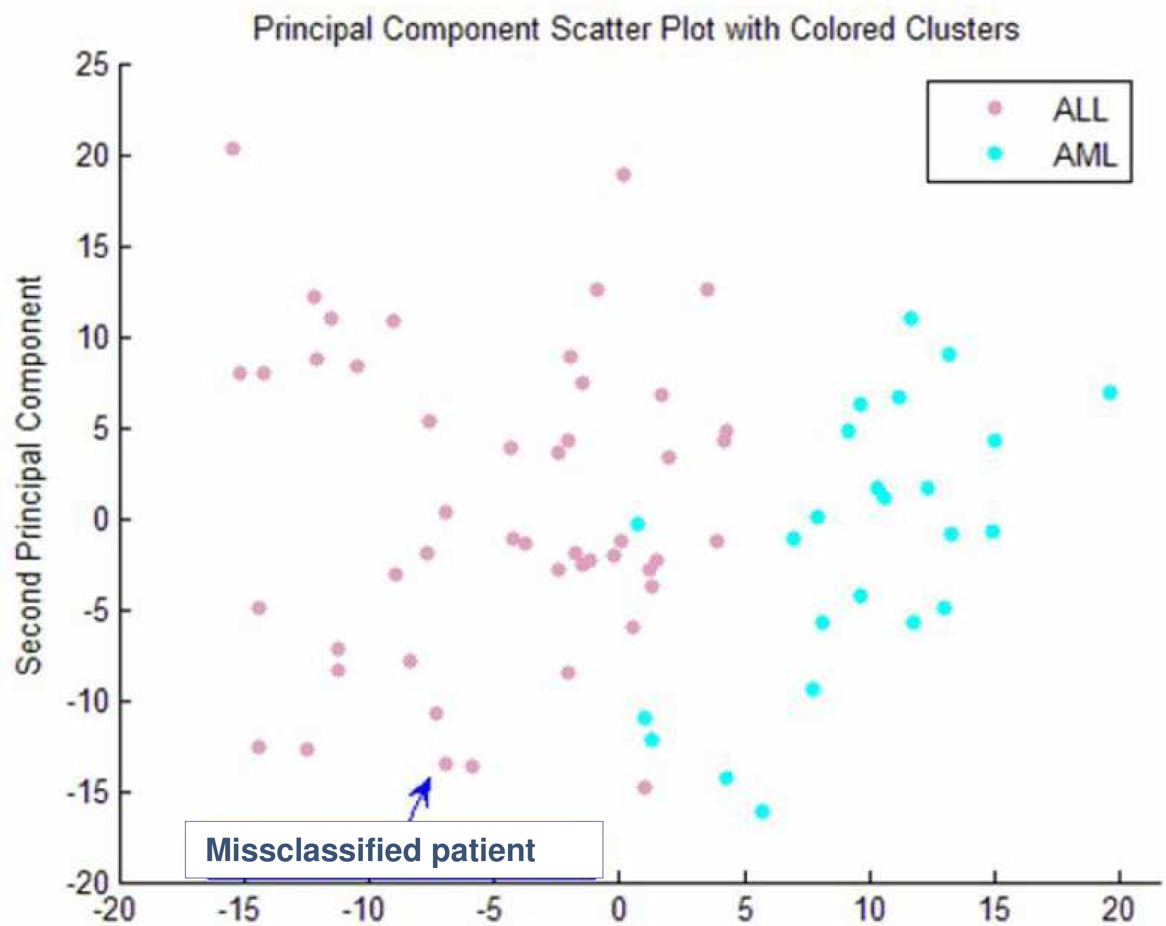
$$F(x_j) = \left|\frac{\mu_j^+ - \mu_j^-}{\sigma_j^+ + \sigma_j^-}\right|$$

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



- Different techniques agree on the miss-classified patient!

Gene expression analysis

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

Dataset	chunk	% of train	genes	% of genes
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***

Research directions

- Is it possible to find an optimal strategy for subset selection?
  - How far (accuracy/computational complexity) is it from the proposed incremental one?
- Is it possible to provide prior knowledge, in generalized eigenvalues classification, analytically rather then with training points?
- Can linear algebra algorithms for large sparse matrices enhance algorithm performance?

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.

High Performance Computing and Networking Institute
National Research Council, Italy

Incremental Classification with Generalized Eigenvalues

Mario Rosario Guarracino

September 17, 2007

Consiglio Nazionale delle Ricerche