CONSTRAINED WEIGHTED FEATURE SELECTION

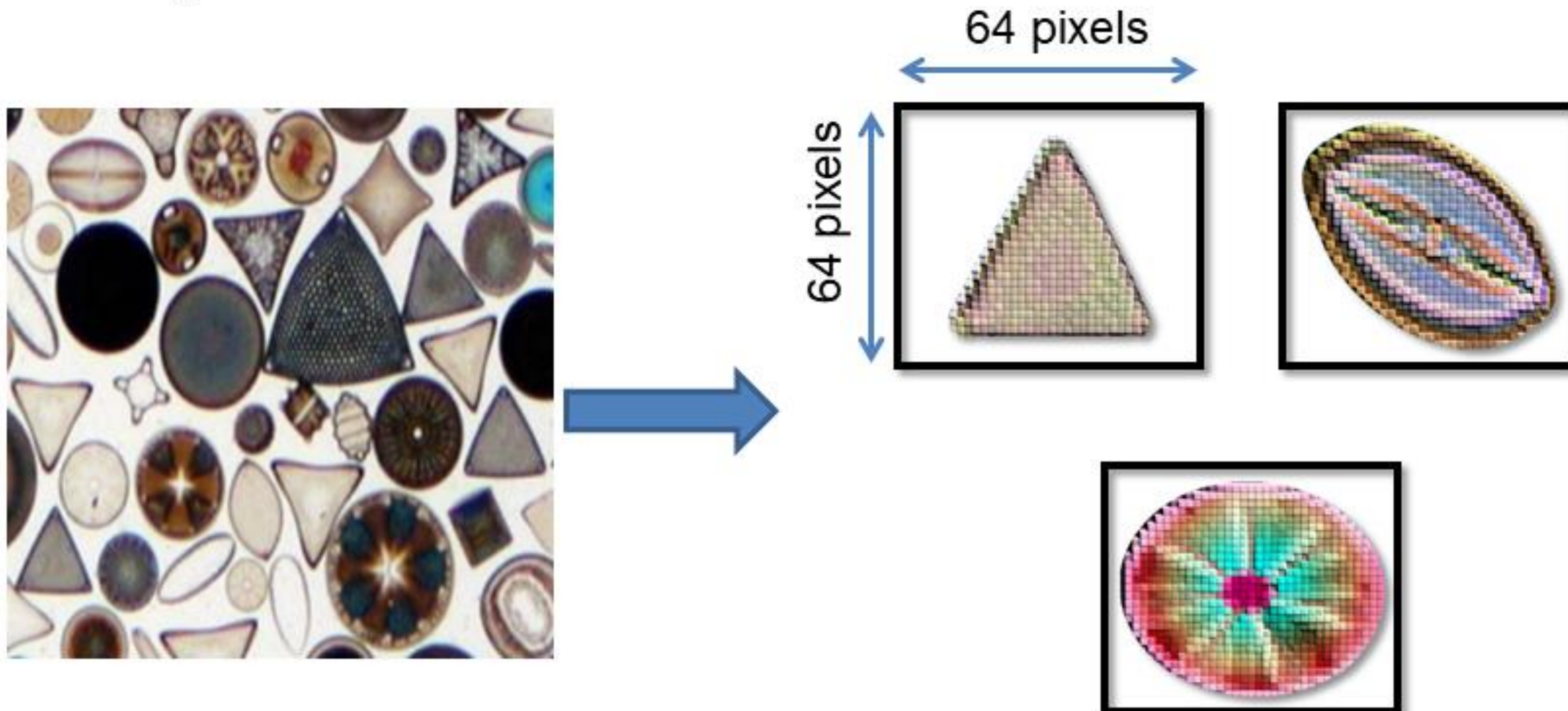
NASCA 2018: Numerical Analysis and Scientific Computation with Application Conference, 2018

Denis Hamad
Samah Hijazi
Mariam Kalakech
Ali Kalakech

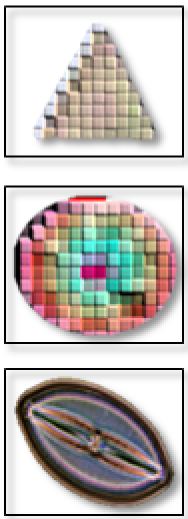
ulco UNIVERSITÉ DU LITTORAL CÔTE D'OPALE
LISIC Laboratoire d'Informatique Signal & Image de la Côte d'Opale
Région Hauts-de-France
CNRS
المجلس الوطني للبحوث العلمية
National Council for Scientific Research
GRAISyHM
Groupement de Recherche en Automatisation Intégrée et Systèmes Hommes-Machines
Université Libanaise
Ecole Doctorale Sciences et Technologie

Context, Motivation & Problem

- In many machine learning applications the data used is provided with a very large number of features.
- Where only few are relevant and not redundant.



Context, Motivation & Problem



f_r

$[x_{11}, \dots, x_{1r}, \dots, x_{1d}]$

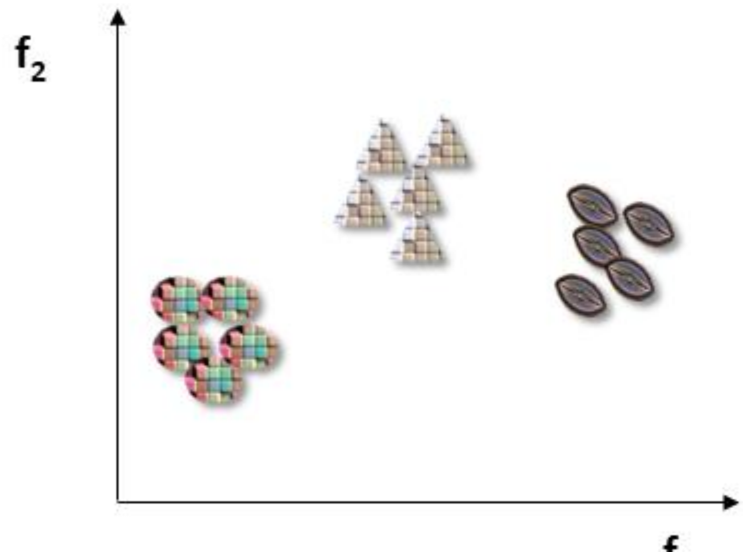
$[x_{n1}, \dots, x_{nr}, \dots, x_{nd}]$ $\boldsymbol{x}_n$ $X(N \times d)$

$[x_{N1}, \dots, x_{Nr}, \dots, x_{Nd}]$

- Each object $\boldsymbol{x}_n (n = 1 \dots, N)$ is characterized by a large number of features f_r $(r =, 1 \dots, d)$.
- x_{nr} represents the value of the n-th data object on the r-th feature.
- The performance of machine learning algorithms might be degraded when applied on such high dimensional data.

Solution: Feature Selection

Coordinates of points represented over irrelevant feature f_2

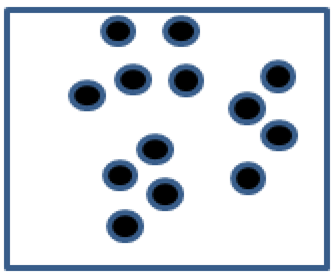
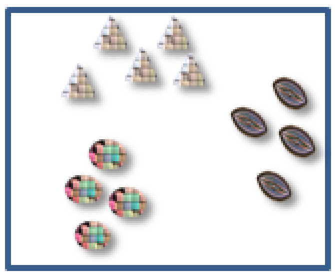
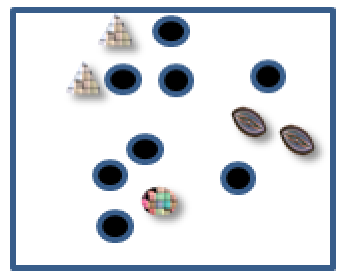
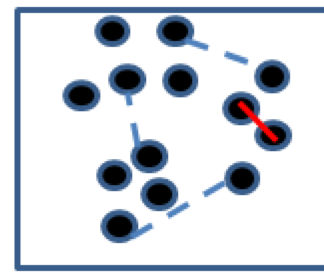


Representation of data points in the original space

Coordinates of points represented over relevant feature f_1

Feature Selection:

Learning Contexts

Unsupervised Learning	**Supervised Learning**	**Semi-supervised Learning**	**Constrained Learning**
			
We know nothing about the data labels	We know the data labels very well	We know something about data labels	We have pairwise constraints on data
• ***Laplacian Score*** • ***Variance Score***	• ***Fisher Score*** • ***Relief Algorithm***	• ***Semi-supervised Laplacian Score*** • ***Semi-supervised logistic I-Relief***	• ***Simba-Sc*** ***Relief-Sc***

Outline

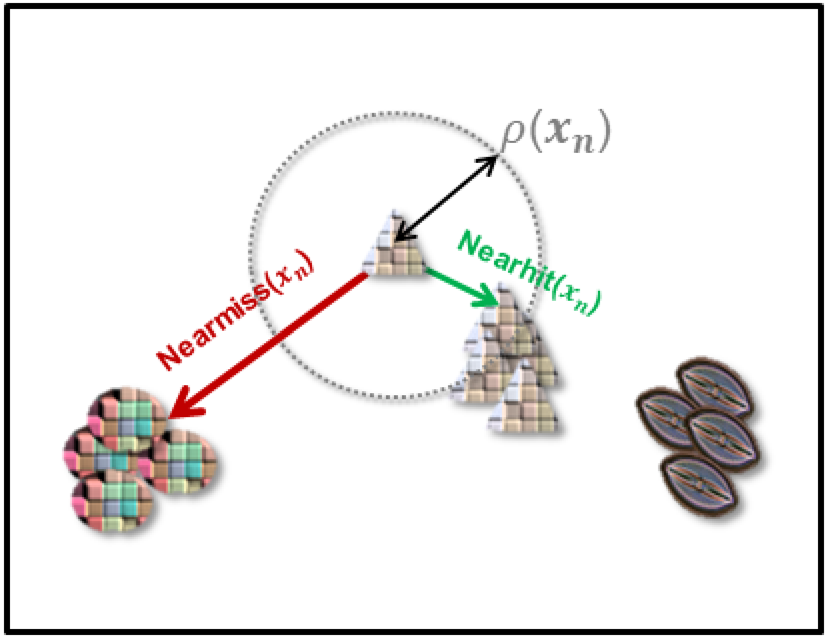
- Hypothesis-Margin: Relief-Sc (Relief with Side Constraints)
- Algorithmic Comparison
- Experimental results on Relief-Sc
- Constraint Selection
- Constraint propagation
- Conclusion & Future Work

Hypothesis-Margin

Supervised Context

Kira K, Rendell LA (1992)

- **Nearmiss** of an instance x_n is the nearest sample to x_n with a different label
- **Nearhit** of an instance x_n is the nearest sample to x_n with the same label



- **Hypothesis Margin** of x_n denoted $\rho(x_n)$ is the **difference** between the distance to its **nearmiss** and the distance to its **nearhit**.

- **Hypothesis Margin** is the largest distance an instance of the dataset can travel without altering the labeling of instances.

Gilad-Bachrach R., Navot A., Tishby (N. 2004)

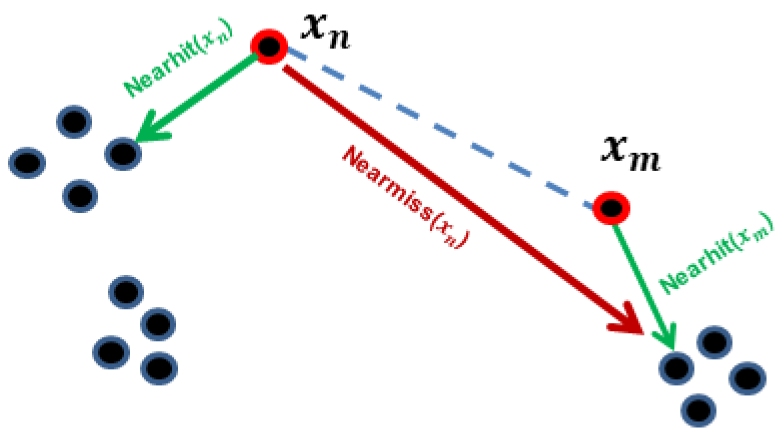
$$\rho(x_n) = \text{diff}\,(x_n, \mathit{Nearmiss}(x_n)) - \text{diff}\,(x_n, \mathit{Nearhit}(x_n))$$

Hypothesis-Margin

Constrained Context

Yang M., Song J. (2010)

- **Nearmiss** of an instance x_n is the nearest sample to x_m
- **Nearhit** of an instance x_n is the nearest sample to x_n



$$\rho(x_n, x_m) = \text{diff}\ (\ x_n\ ,\ Nearhit(x_m)) - \text{diff}\ (\ x_n\ ,\ Nearhit(x_n))$$

Constrained Weighted Feature Selection

- Weighted Hypothesis Margin of $(\boldsymbol{x}_n, \boldsymbol{x}_m)$

$$\rho(\boldsymbol{\omega}, (\boldsymbol{x}_n, \boldsymbol{x}_m)) = \boldsymbol{\omega}^T \left[\text{diff}\ (\boldsymbol{x}_n, \boldsymbol{Nearhit}(\boldsymbol{x}_m)) - \text{diff}\ (\boldsymbol{x}_n, \boldsymbol{Nearhit}(\boldsymbol{x}_m))\right]$$

- Overall Weighted margin

$$\rho = \boldsymbol{\omega}^T \underbrace{\sum_{(x_n, x_m) \in C} \left[\text{diff}\ (\boldsymbol{x}_n, \boldsymbol{Nearhit}(\boldsymbol{x}_m)) - \text{diff}\ (\boldsymbol{x}_n, \boldsymbol{Nearhit}(\boldsymbol{x}_m))\right]}_{\boldsymbol{z}}$$

Sun Y. and Li J. (2006)

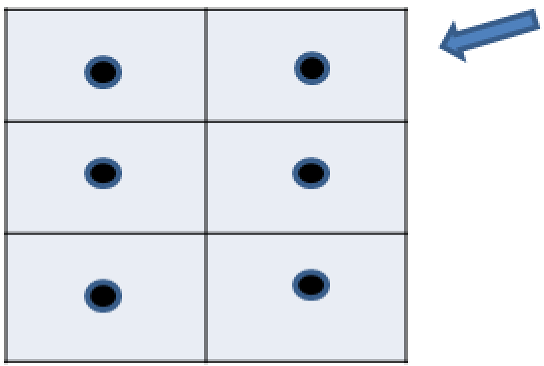
- Problem Formulation:

  — We want to find the weight vector $\boldsymbol{\omega}$ that maximizes the overall margin
  — Then the Problem is defined as

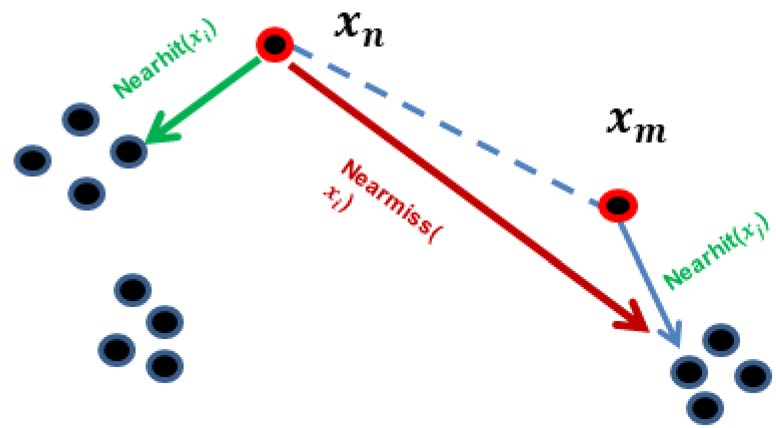
$$\max_{\omega} \ \boldsymbol{\omega}^T \boldsymbol{z} \qquad \text{s.t. } \|\boldsymbol{\omega}\|^2 = 1 \ and \ \boldsymbol{\omega} \geq 0$$

where $\boldsymbol{\omega} = (\boldsymbol{\omega}_1, \ldots, \boldsymbol{\omega}_r, \ldots, \boldsymbol{\omega}_d)$

Relief-Sc
(Relief with Side Constraints)



Cannot-link Constraints Set

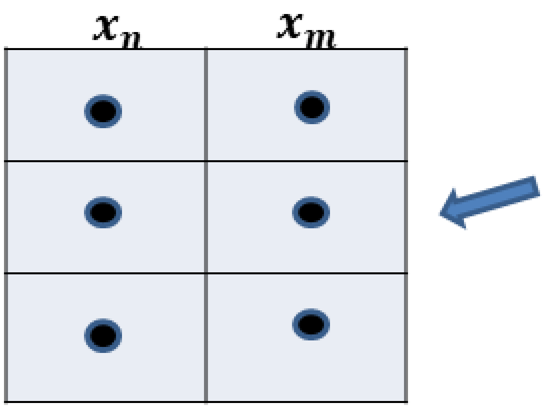


$$
\begin{array}{c|ccccc|}
 & f_1 & & f_r & & f_d \\
x_n & x_{11} & \cdots & x_{1r} & \cdots & x_{1d} \\
 & \cdots & \cdots & \cdots & \cdots & \cdots \\
x_m & x_{n1} & \cdots & x_{nr} & \cdots & x_{nr} \\
 & \cdots & \cdots & \cdots & \cdots & \cdots \\
 & x_{N1} & \cdots & x_{Nr} & \cdots & x_{Nd}
\end{array}
$$

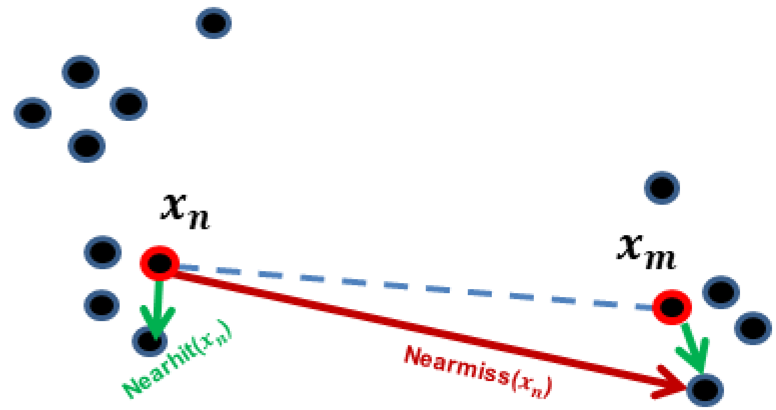
z = (0 , 0 … 0)

$$z_r = z_r + |x_{nr} - Nearhit(x_{mr})| - |\, x_{nr} - Nearhit(x_{nr})|$$

Relief-Sc
(Relief with Side Constraints)



Cannot-link Constraints Set



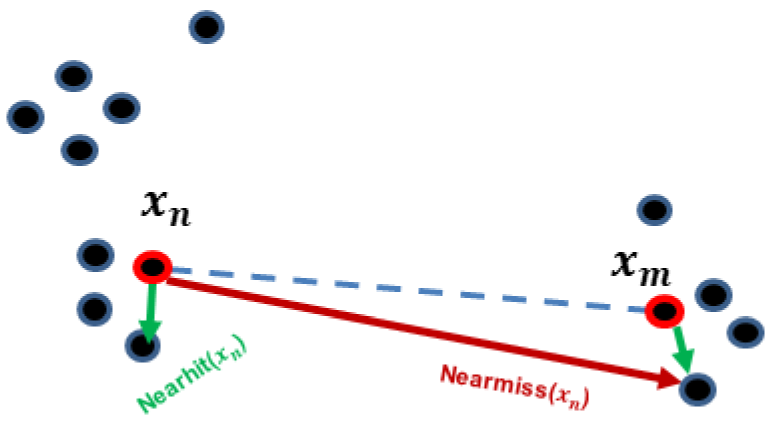
$$z_r = z_r + |x_{nr} - Nearhit(x_{mr})| - |\, x_{nr} - Nearhit(x_{nr})|$$

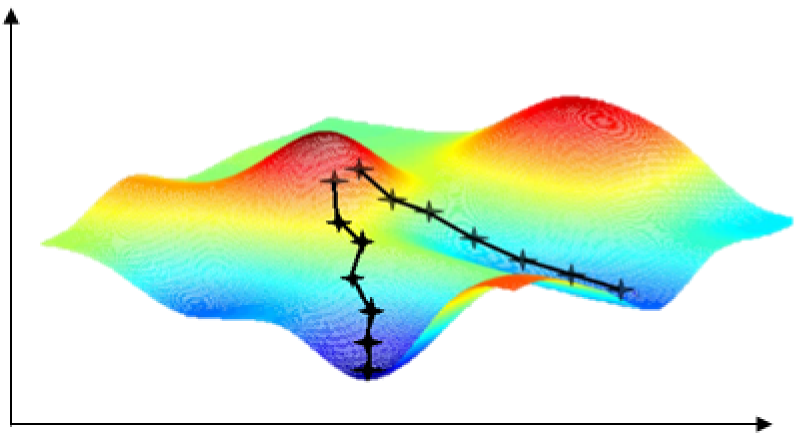
$$\mathbf{z}^+ = [\max(z_1, 0), \ldots, \max(z_d, 0)]^T$$

$$\boldsymbol{\omega} = \frac{\mathbf{z}^+}{\|\mathbf{z}^+\|}$$

f_1	f_r	f_d
$[\omega_1$	ω_r	$\omega_d]$

Algorithmic comparison:
Simba-Sc





Yang M., Song J. (2010)

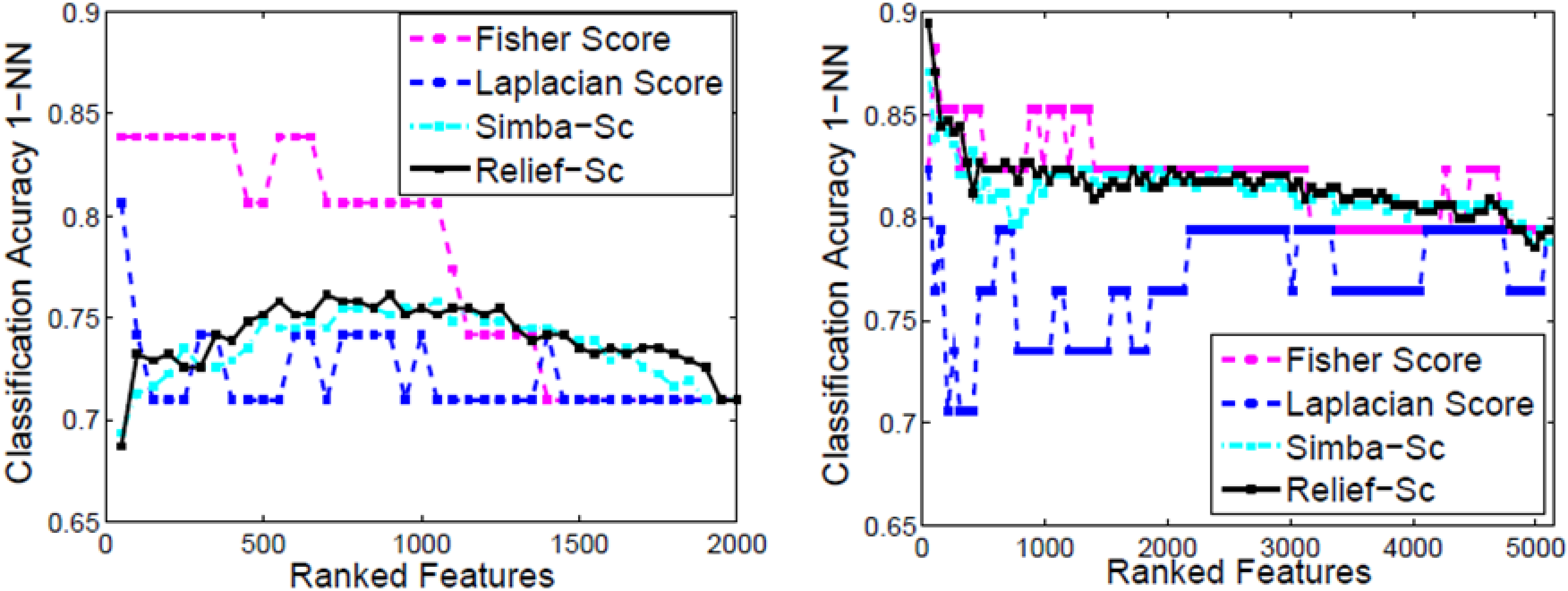
$$\omega_r = \omega_r + \frac{1}{2}\left[\frac{(x_{nr} - Nearhit(x_{mr}))^2}{\|x_n - Nearhit(x_m)\|_\omega} - \frac{(x_{nr} - Nearhit(x_{nr}))^2}{\|x_n - Nearhit(x_n)\|_\omega}\right]\omega_r$$

$$\boldsymbol{\omega} = \frac{\boldsymbol{\omega}}{\|\boldsymbol{\omega}\|_\infty}$$

f_1	f_r	f_d
$[\omega_1$	ω_r	$\omega_d]$

Experimental Results:

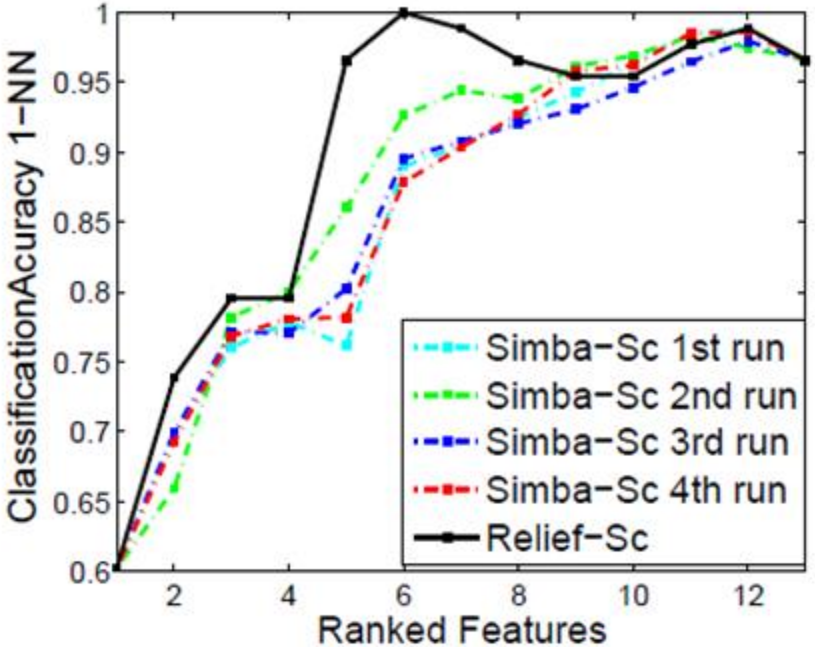
Classification Accuracy of Different Feature Selection Algorithms



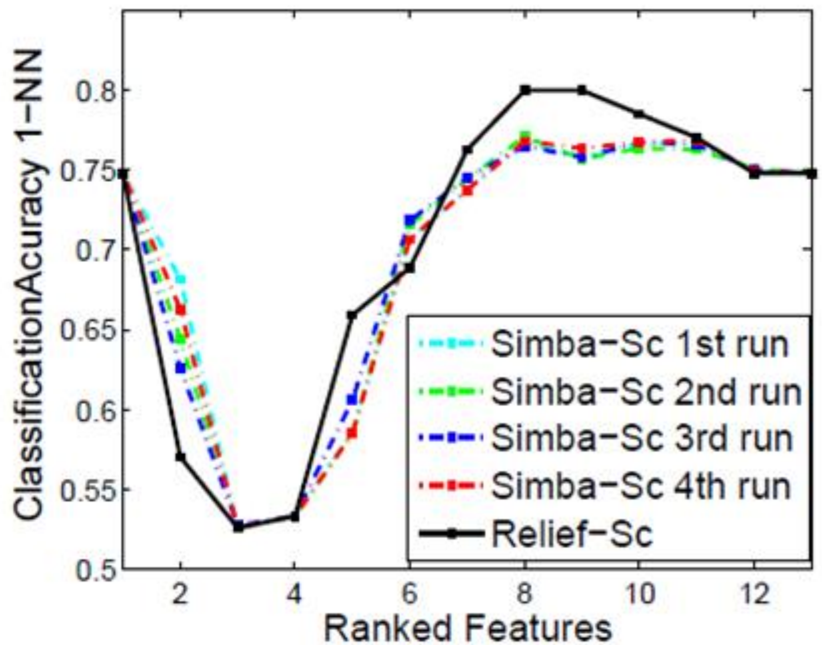
(a) ColonCancer (b) Leukemia

Experimental Results:

Solutions Comparison of Relief-Sc and Simba-Sc

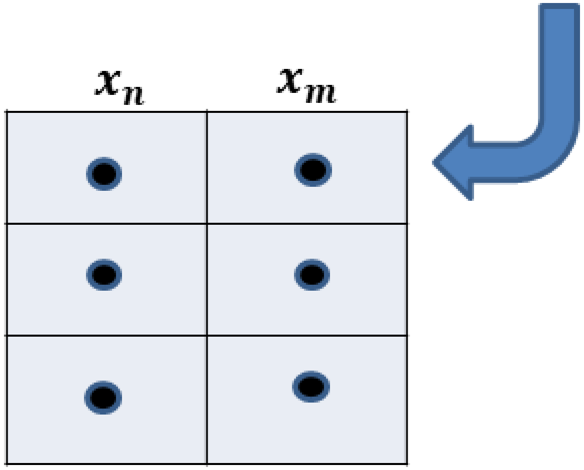


(a) Wine



(b) Heart

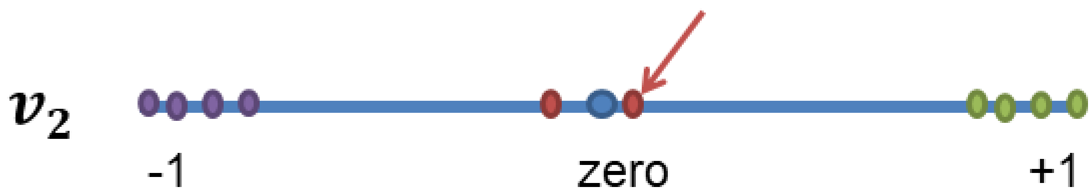
Selection of Cannot-link Constraint Set



Cannot-link Constraints Set

1. Find the Laplacian matrix L of the data
2. Find eigenvectors V and eigenvalues λ of L=D-S

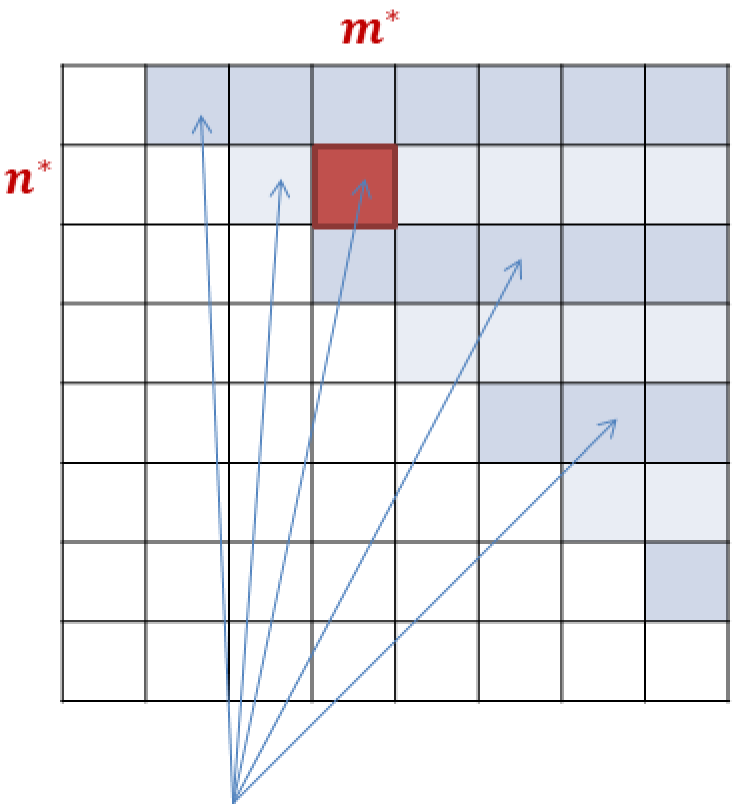
3. Find the point of minimum value on the second eigenvector v_2



4. Find the data couple that can most change the position of the minimum value $v_{2(i^*)}$ on $\boldsymbol{v_2}$ such that:

$$(x_{n^*}, x_{m^*}) = argmax_{n,m \in \{1 \ldots N\}} \left| \frac{dv_{2(i^*)}}{ds_{nm}} \right|$$

Selection of Cannot-link Constraint Set



Sensitivities of (x_n, x_m)

We find the sensitivity of each couple and then The maximum $(\boldsymbol{x}_{n^*}, \boldsymbol{x}_{m^*})$ among all.

4. Thus, we calculate the sensitivity of $v_{2(i^*)}$ to each couple in the dataset and store them in a matrix.

Now, we find the data having the highest sensitivity

$$(\boldsymbol{x}_{n^*}, \boldsymbol{x}_{m^*}) = argmax_{n,m \in \{1 \dots N\}} \left| \frac{dv_{2(i^*)}}{ds_{nm}} \right|$$

Where

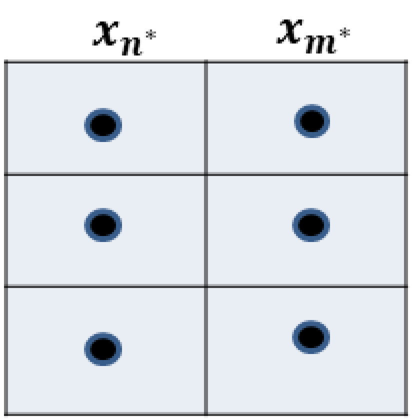
$$\left| \frac{dv_{2(i^*)}}{ds_{nm}} \right| = \left| \sum_{p>2}^{N} \frac{\boldsymbol{v}_2^T [\partial \boldsymbol{L} / \partial s_{nm}] \boldsymbol{v}_p}{\lambda_2 - \lambda_p} v_{p(i^*)} \right|$$

$$= \left| \sum_{p>2}^{N} \frac{\boldsymbol{v}_2^T [(\boldsymbol{e}_n - \boldsymbol{e}_m)(\boldsymbol{e}_n - \boldsymbol{e}_m)^T] \boldsymbol{v}_p}{\lambda_2 - \lambda_p} v_{p(i^*)} \right|$$

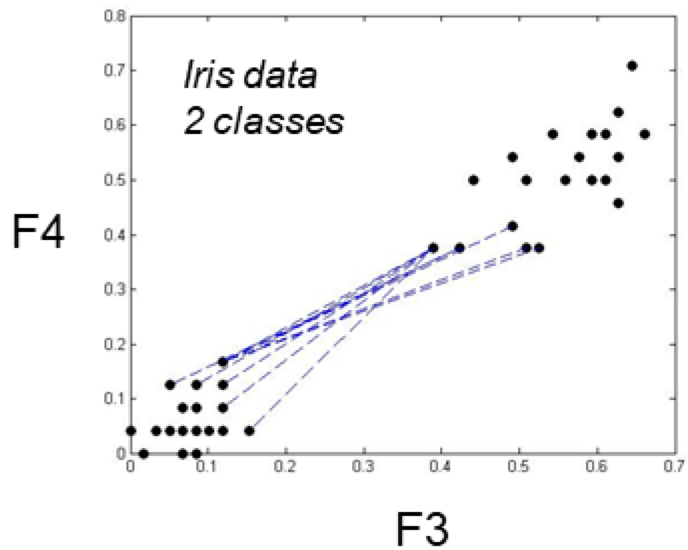
Selection of Cannot-link Constraint Set

5. After finding the indexes (n^*, m^*), We Actively query for a constraint on this particular couple.

6. Thus, obtaining the constraints of highest utility.



Cannot-link Constraints Set

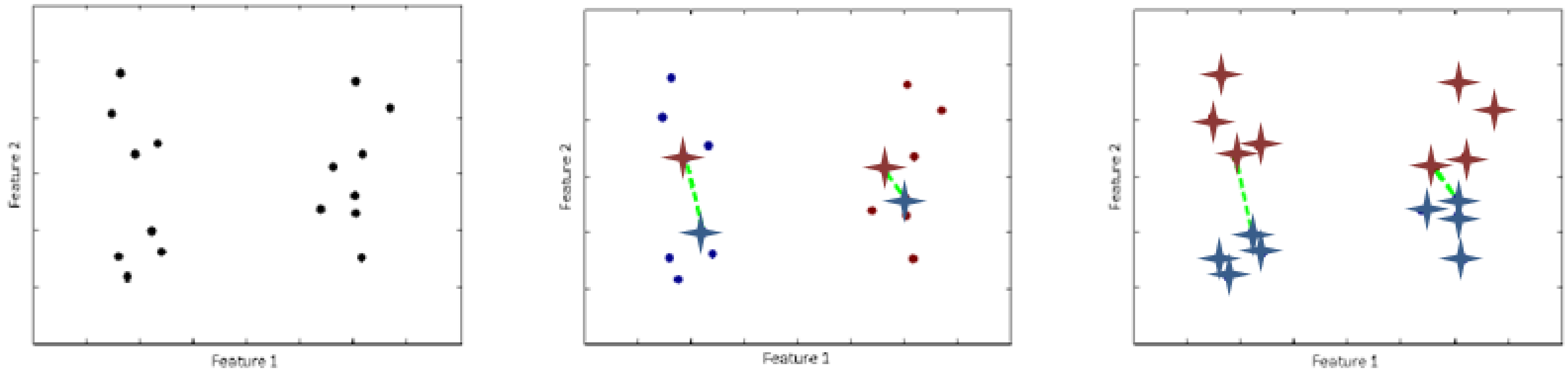


However, we can't ask the user a lot of questions, so the number of constraints we can obtain is limited

Constraints Propagation

- Relief family margin-based algorithms need a set of points that is sufficiently big to calculate the margin.
- we query an oracle for constraints
- we might not be able to ask for a large number of queries

Therefore we needed a constraint propagation method



(a) Data (b) Instance-level constraints (c) Propagated information

Constraints Propagation

- We initialize the constraints matrix Q_{nm} (from previous constraints selection step) as follows:

$$Q_{nm} = \begin{cases} 1, & \text{if } (x_n, x_m) \in \mathcal{C} \\ 0, & \text{otherwise} \end{cases}$$

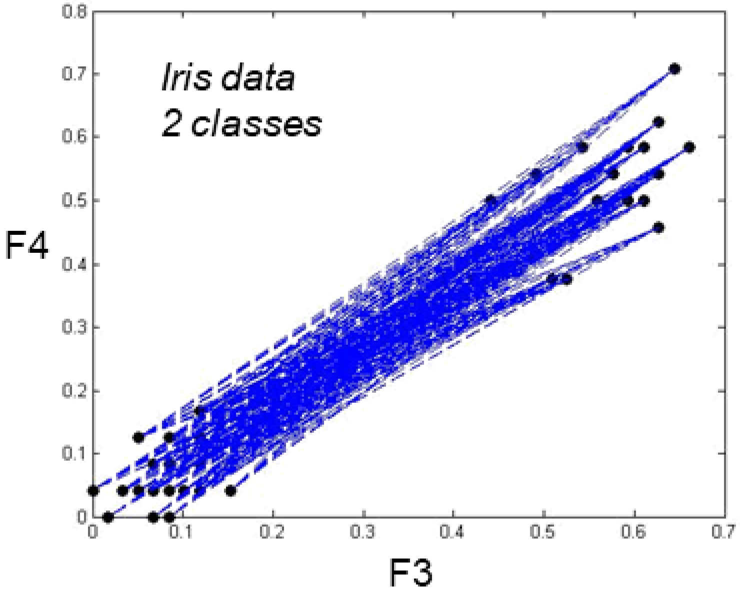
- We calculate the normalized neighborhood similarity graph P_{nm}

$$P_{nm} = \begin{cases} e^{-\frac{\|x_n - x_m\|^2}{2\sigma^2}}, & \text{if } \quad x_m \in k\text{-NN}(x_n) \\ 0, & \text{otherwise} \end{cases}$$

- Finally we use the following matrix completion equation to propagate our constraint on P_{nm} obtaining G^*

$$G^* = (1-a)^2 (I - aP)^{-1} Q (I - aP)^{-1}$$

Where I is the identity matrix and a is a regularization parameter



Conclusion

- We proposed Relief-Sc, a **weighted feature selection** algorithm that works in a **constrained** environment.
- It is said to find a **unique** relevant feature subset in a **closed-form**.
- We are currently working on the experimental results of Constraints selection and Propagation, we expect to obtain better feature selection with a minimum number of queried constraints.

Acknowledgments

- This work is supported by a scholarship granted by the University of the Littoral Opal Coast (ULCO) and (AUF) in France together with the National Council For Scientific Research in Lebanon (CNRS-L) as a part of ARCUS E2D2 project.

References

- Gilad-Bachrach R., Navot A., Tishby N. (2004), "Margin based feature selection—theory and algorithms », 21st Int. Conf. on Machine Learning, Canada, pp. 43–50.
- Kira K, Rendell L.A. (1992) A practical approach to feature selection. In: Proceedings of the ninth int. workshop on Machine learning, pp 249-256.
- Lu Z., Ip H.H. (2010) Constrained spectral clustering via exhaustive and efficient constraint propagation. In: European Conf. on Computer Vision, Springer, pp 1-14.
- Sun Y. and Li J. (2006) "Iterative RELIEF for Feature Weighting," Proc. 23rd Int'l Conf. Machine Learning, pp. 913-920.
- Wauthier F.L., Jojic N., Jordan M.I. (2012) Active spectral clustering via iterative uncertainty reduction. In: Proceedings of the 18th ACM SIGKDD int. conf. on Knowledge discovery and data mining, ACM, pp 1339-1347.
- Yang M., Song J. (2010) A novel hypothesis-margin based approach for feature selection with side pairwise constraints. Neurocomputing, 73(16):2859-2872.