Learning Maximum Excluding Ellipsoids In Unbalanced Scenarios with Theoretical Guarantees

Guillaume Metzler - PhD Student

Supervised by : M. Sebban, A. Habrard and E. Fromont

Summer School 2017
Big Data & Business Methods
Technologies and Innovation

LABORATOIRE HUBERT CURIEN
UMR • CNRS • 5516 • SAINT-ETIENNE
UNIVERSITÉ DE LYON
UNIVERSITÉ JEAN MONNET SAINT-ÉTIENNE

Outline

- Origin of the idea
- Presentation of the algorithm
- Theoretical Guarantees
- Applications & Experiments

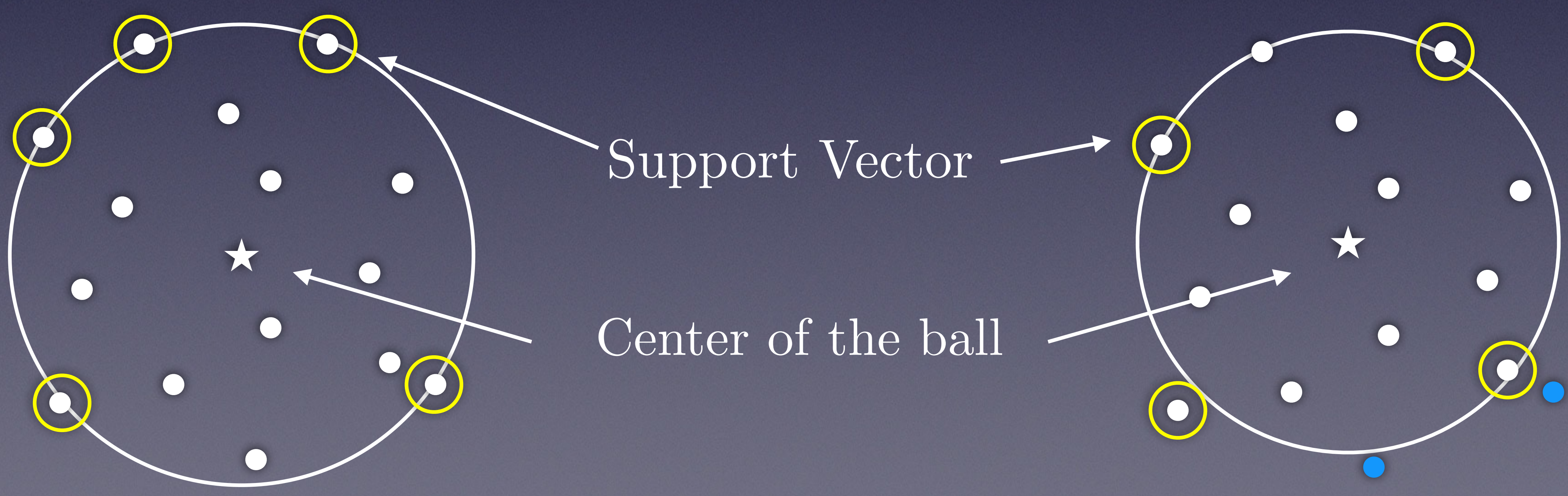
The Minimum Including Ball Problem

Given a set of n **unlabelled** points, find the center $\mathbf{c}$ and the smallest radius R of the ball that includes the data [Tax & Duin (2004)]

$$\min_{\mathbf{c},R} \quad R^2 + \frac{\mu}{n}\sum_{i=1}^{n}\xi_i$$

$$s.t. \quad \|\mathbf{x}_i - \mathbf{c}\| \leq R^2 + \xi_i, \; \forall i = 1, ..., n$$

Hard Inclusion ($\xi_i = 0$) Soft Inclusion ($\xi_i \geq 0$)

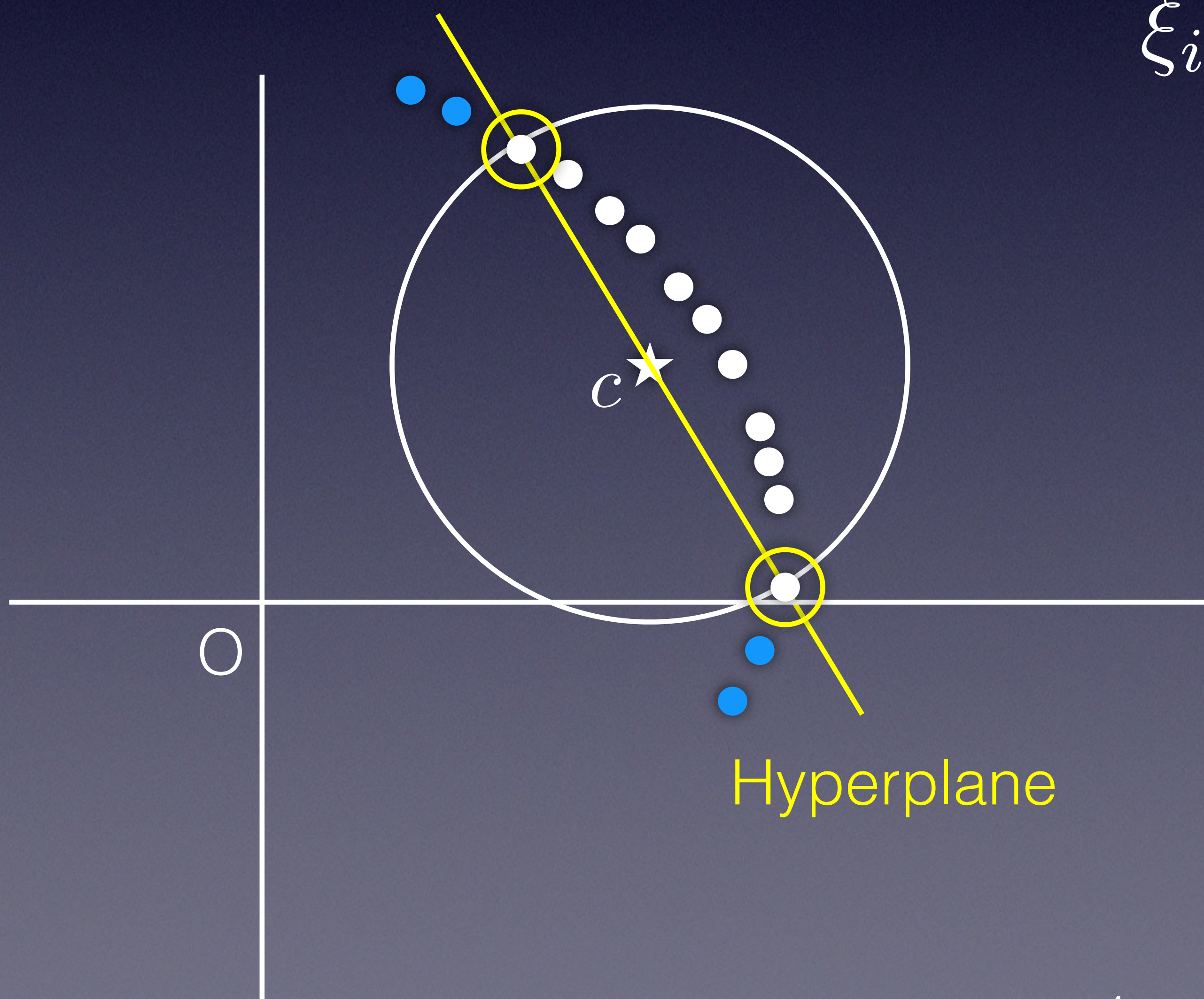


MIB and One class SVM

$$\min_{\mathbf{c},\xi,\rho} \quad \frac{1}{2}\|\mathbf{c}\|^2 + \frac{1}{\nu n}\sum_{i=1}^{n}\xi_i - \rho - \frac{1}{2}\|\mathbf{x}_i\|$$

$$s.t. \qquad \mathbf{c}^T\mathbf{x}_i \geq \rho + \frac{1}{2}\|\mathbf{x}_i\| - \xi_i$$

$$\xi_i \geq 0$$

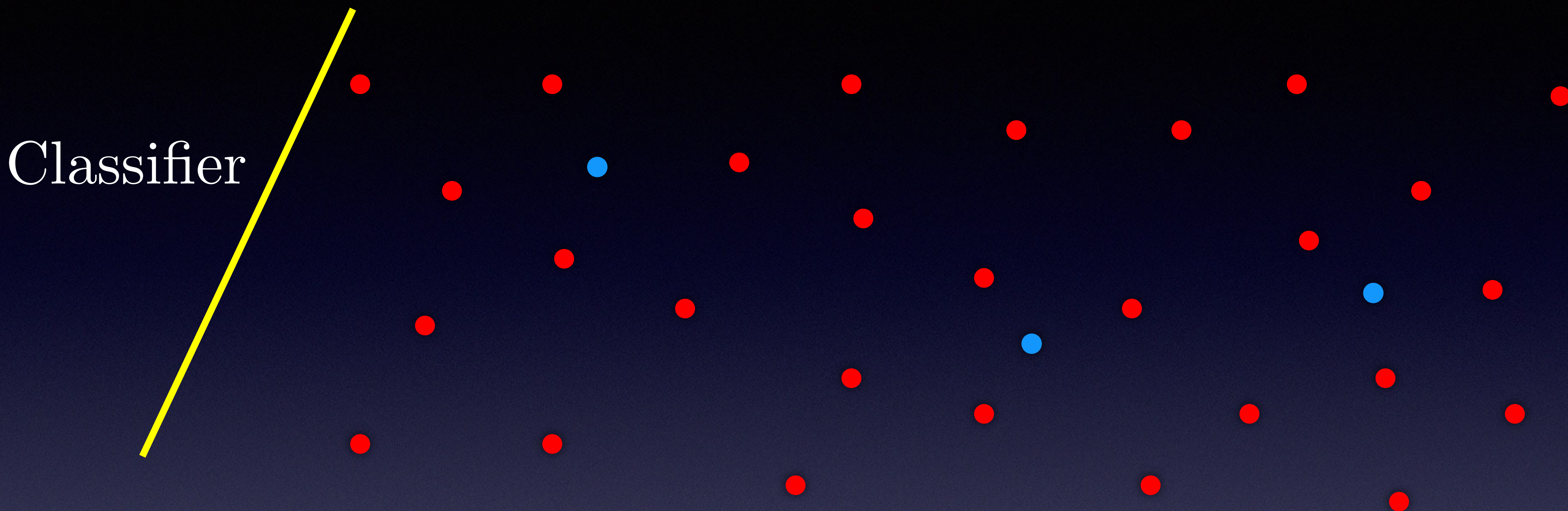


Belonging to the ball

$\Updownarrow$

Being above the hyperplane

Anomaly / Fraud detection: towards a Maximum Excluding Ball problem

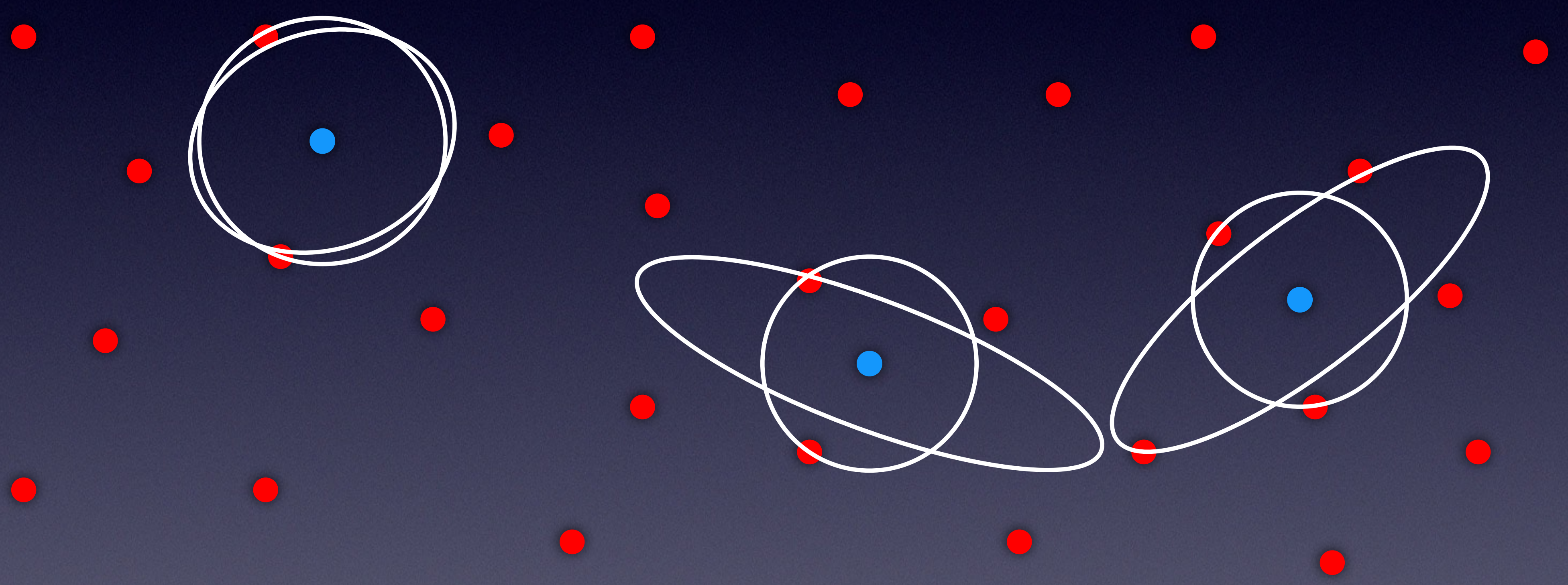


Maximizing the accuracy is inappropriate.
More relevant criteria:

$$Precision = \frac{TP}{TP + FP} \qquad Recall = \frac{TP}{TP + FN}$$

$$F_\beta = \frac{(1+\beta^2) Precision \cdot Recall}{\beta^2 \cdot Precision \cdot Recall}$$

A Metric Learning-based approach



From excluding balls to learned excluding ellipsoids

$\Rightarrow$ The Maximum Excluding Ellipsoid (ME^2) algorithm

Key Properties of ME^2

- center of ellipsoids are no more learned (positive data)
- negative examples are used to learn ellipsoids
- use of a Mahalanobis like metric learning approach:

$$\|\mathbf{x} - \mathbf{c}\|_{\mathbf{M}}^2 = (\mathbf{x} - \mathbf{c})^T \mathbf{M} (\mathbf{x} - \mathbf{c})$$

s.t. $\mathbf{M}$ is $\mathbf{PSD}$.
ME^2 comes with a cheap way to get the positive definiteness of $\mathbf{M}$.

The ME^2 Algorithm

Given a set of n negative examples and p positive examples i.i.d. according to a joint distribution $\mathcal{D}$ over $\mathbb{R}^d \times \{-1, +1\}$.
For all positive examples $\mathbf{c}$:

$$
\begin{aligned}
\min_{R,\mathbf{M},\xi} \quad & \frac{1}{n}\sum_{i=1}^{n} \xi_i + \mu(B-R)^2 + \lambda\|\mathbf{M}-\mathbf{I}\|_F^2, \\
s.t. \quad & \|\mathbf{x}_i - \mathbf{c}\|_{\mathbf{M}}^2 \geq R - \xi_i, \quad \forall i = 1, ..., n, \\
& \xi_i \geq 0, \\
& B \geq R \geq 0.
\end{aligned}
$$

R	radius of the ellipsoid	B	bound on ellipsoid's size
$\mathbf{M}$	$d \times d$ matrix	μ	controls ellipsoid's size
ξ	slack variables	λ	controls distortion w.r.t. to a ball

Dual Formulation

$$
\begin{aligned}
\min_{\alpha,\beta,\delta} \quad & \alpha^T \left(\frac{1}{4\lambda}\mathbf{G}' + \frac{1}{4\mu}\mathbf{U}_{d\times d} \right) \alpha + \frac{\beta^2}{4\mu} + \frac{\delta^2}{4\mu} + \\
& \alpha^T \left(diag(\mathbf{G}) - \left(B + \frac{\beta}{2\mu} - \frac{\delta}{2\mu} \right) \mathbf{U}_d \right) + \beta \left(B - \frac{\delta}{2\mu} \right), \\
s.t. \quad & 0 \le \alpha_i \le \frac{1}{n}, \quad \forall i = 1, ..., n, \\
& \beta, \delta \ge 0,
\end{aligned}
$$

where $\mathbf{G}$ is the Gram matrix and $\mathbf{G}'$ is the Hadamard product of $\mathbf{G}$ with itself. $\mathbf{U}_d$ (respectively $\mathbf{U}_{d\times d}$) represents a vector of length d (respectively a matrix of size $d \times d$) where entries are equal to 1.

About the Dual Formulation

- Easier to solve, only depends on the number of positive instances.
- Gives an explicit expression of both Radius R and Similarity $\mathbf{M}$

$$R = \frac{\beta - \delta + 2\mu B - \sum_{i=1}^{n} \alpha_i}{2\mu}$$

$$\mathbf{M} = \mathbf{I} + \frac{1}{2\lambda} \sum_{i=1}^{n} \alpha_i (\mathbf{x}_i - \mathbf{c})(\mathbf{x}_i - \mathbf{c})^T$$

- Last equality shows that $\mathbf{M}$ is **PSD**.

Theoretical Guarantees derived from ME^2

Theoretical Results

Uniform Stability [O.Bousquet & A.Elisseeff (2002)]

Definition

A learning algorithm has a uniform stability in $\frac{\beta}{n}$ respect to a loss function ℓ and a parameter set θ, with β a positive constant if:

$$\forall S,\ \forall i,\ 1 \leq i \leq n,\ \sup_{\mathbf{x}} |\ell(\theta_S, \mathbf{x}) - \ell(\theta_{S^i}, \mathbf{x})| \leq \frac{\beta}{n}.$$

where S^i corresponds to S after the replacement of one example drawn according to $\mathcal{D}$.

Theorem

Let $\delta > 0$ and $n > 1$. For any algorithm with uniform stability β/n, using a loss function bounded by K, with probability $1 - \delta$ over the random draw of S:

$$L(\theta_S) \leq \hat{L}_S(\theta_S) + \frac{2\beta}{n} + (4\beta + K)\sqrt{\frac{\ln 1/\delta}{2n}},$$

where $L(\cdot)$ is the true risk and $\hat{L}_S(\cdot)$ its empirical estimate over S.

Hinge Loss Version of ME^2

Using a hinge loss $\ell(\mathbf{M}, R, \mathbf{x}) = \frac{1}{n}[R - \|\mathbf{x}_i - \mathbf{c}\|^2_{\mathbf{M}}]_+$ the problem can be rewritten as follow:

$$\begin{aligned} \underset{R,\mathbf{M}}{min} \quad & \sum_{i=1}^{n} \ell(R, \mathbf{M}, \mathbf{x}_i) + \mu(B - R)^2 + \lambda\|\mathbf{M} - \mathbf{I}\|^2_F, \\ s.t. \quad & B \geq R \geq 0. \end{aligned}$$

Generalization Guarantee on ME^2

Theorem

Let $\delta > 0$ and $n > 1$. There exists a constant $\kappa > 0$, such that with probability at least $1-\delta$ over the random draw over S, we have for any $(\mathbf{M}, R)$ solution of our optimization problem:

$$L(\mathbf{M}, R) \leq \hat{L}_S(\mathbf{M}, R) + \frac{4\max(1, 4B^2)}{n\kappa\min(\mu, \lambda)} + \left(\frac{8\max(1, 4B^2)}{\kappa\min(\mu, \lambda)} + B + 4B^2\sqrt{\frac{\mu B^2}{\lambda} + d}\right)\sqrt{\frac{\ln 1/\delta}{2n}}$$

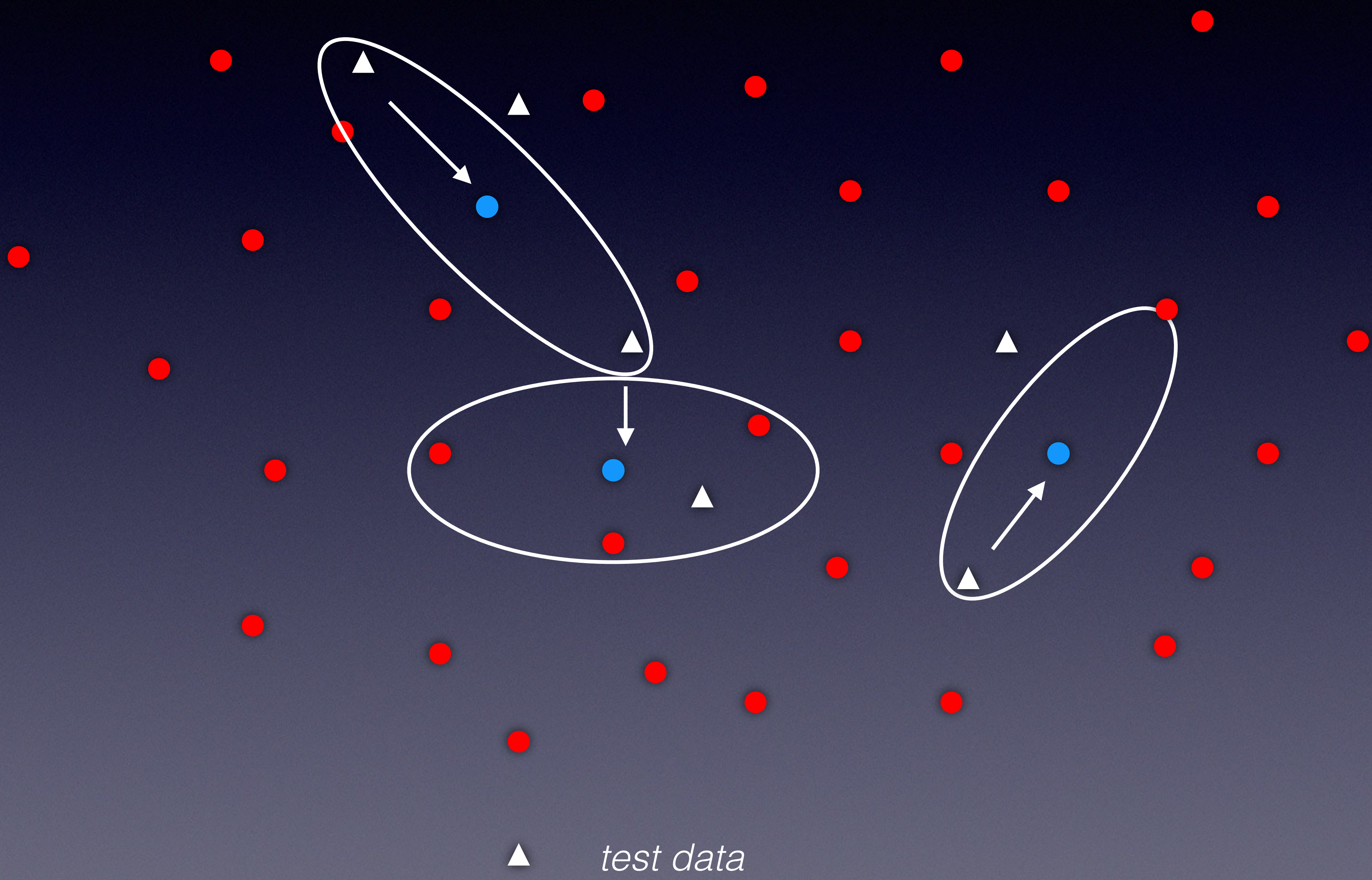
with $\beta = \dfrac{2}{\kappa min(\mu, \lambda)}(max(1, 4B^2))^2$

and $K = B + 4B^2\sqrt{\dfrac{\mu B^2}{\lambda} + d}.$

Experimental Results

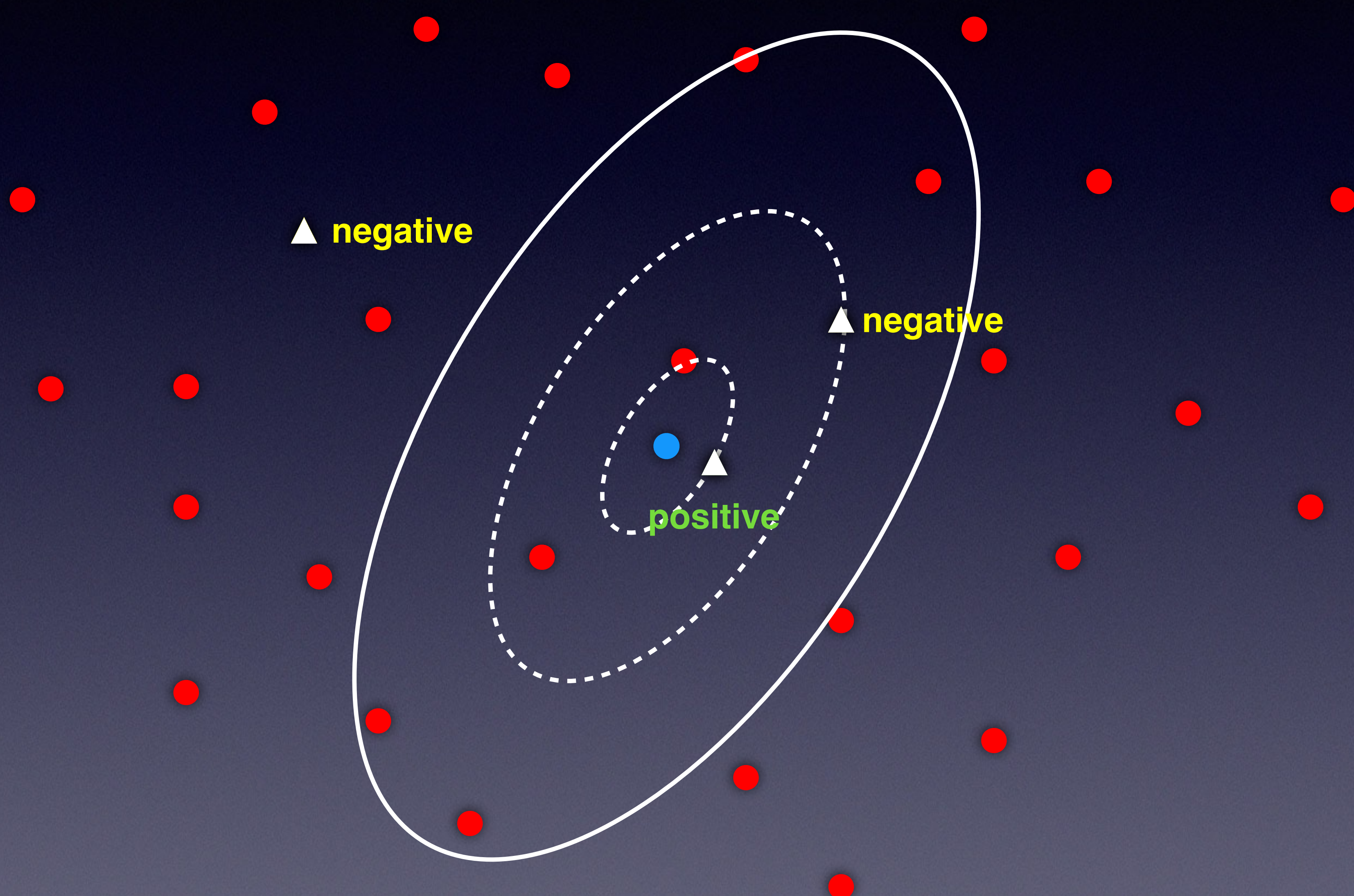
A neighborhood based decision rule

At test time, a data is assigned to its nearest center



A neighborhood based decision rule

Label prediction



ME^2 tends to maximize the F-Measure

Experimental Comparison

- Decision Tree
- Decision Tree with sampling methods:
    1. Oversampling
    2. Undersampling
    3. Both
- Random Forest
- SVM with Linear Kernel
- SVM with Gaussian Kernel

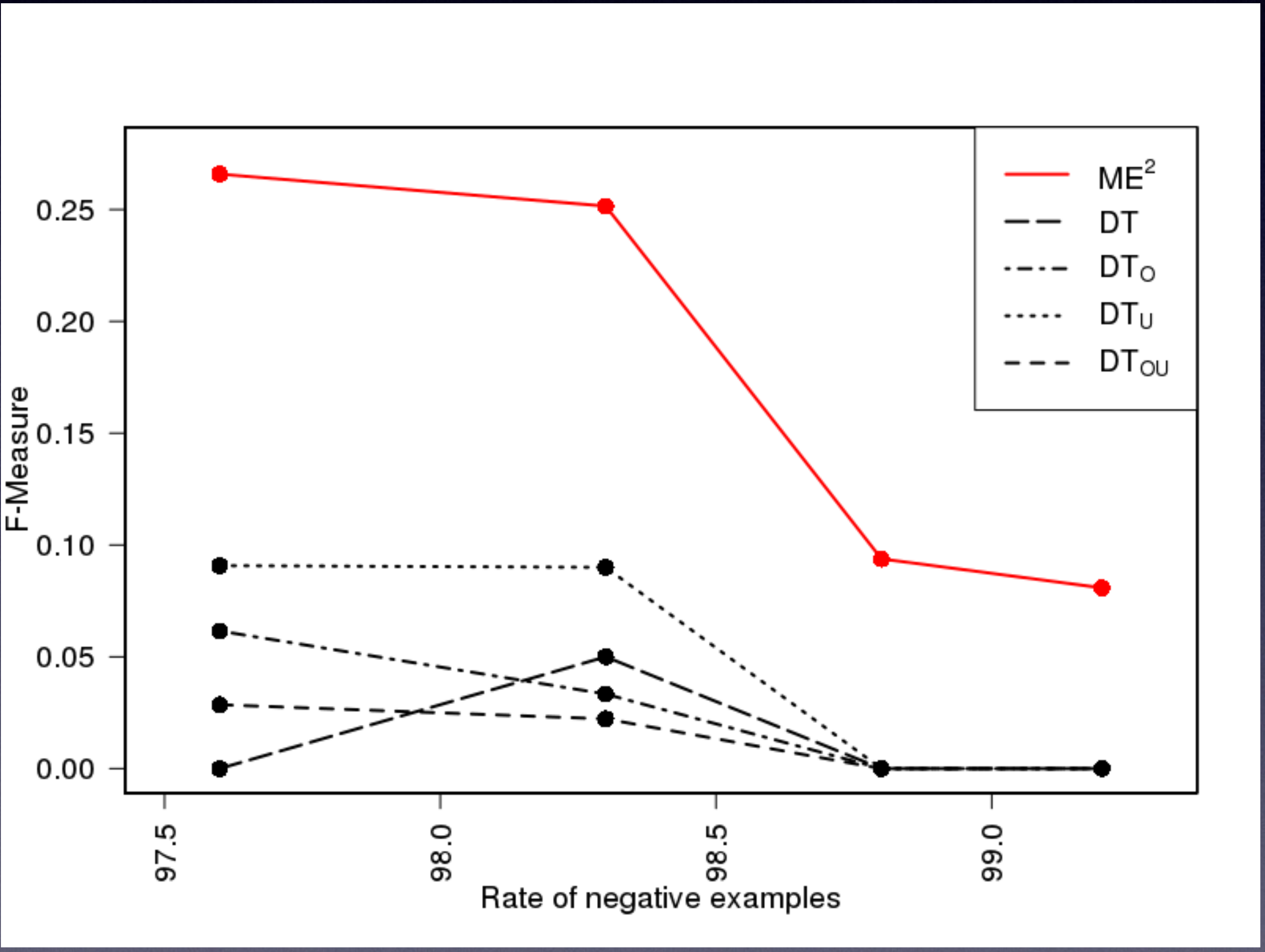
Experimental Comparison

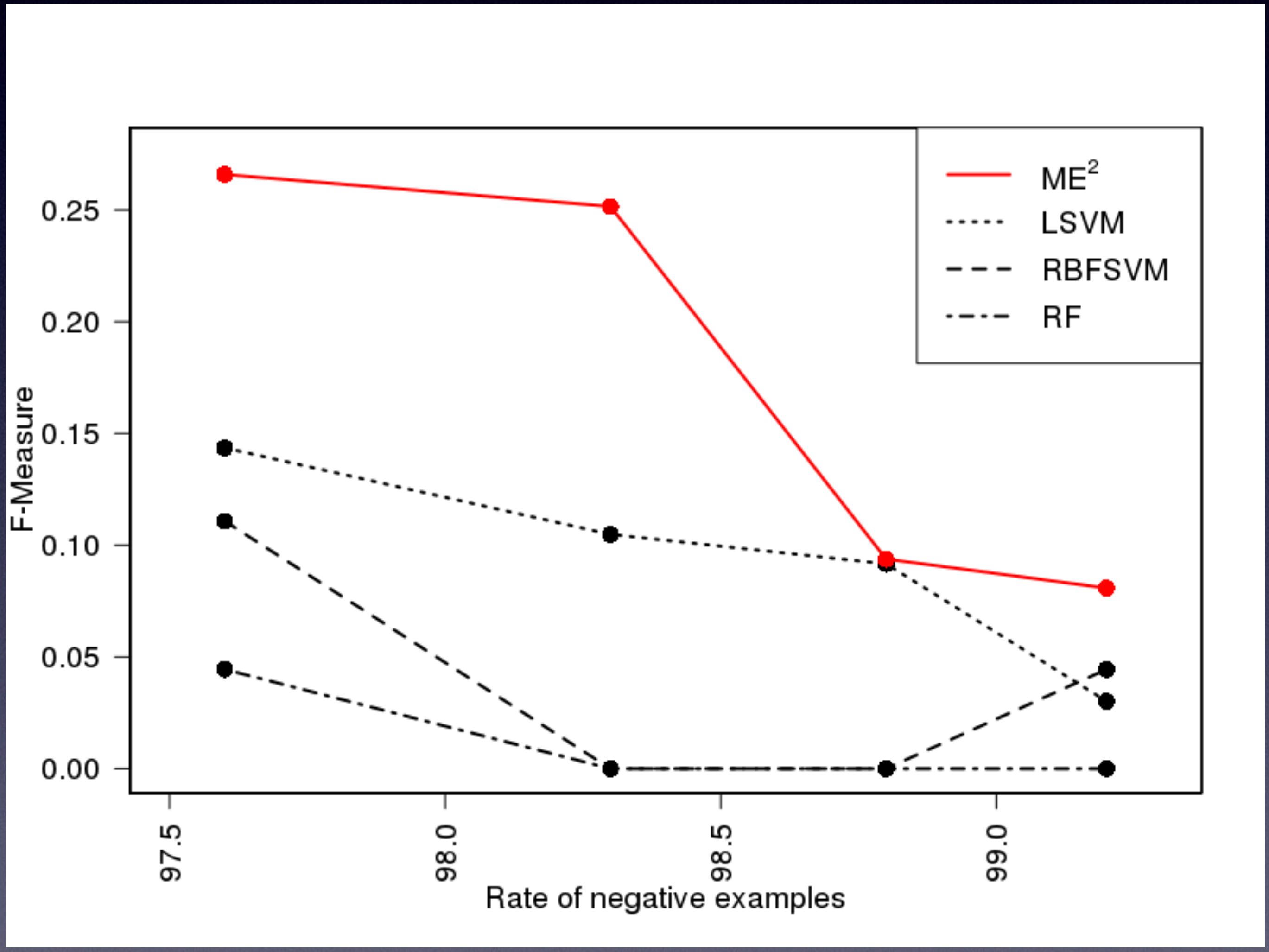
Algorithm	Abalone	Wine	Abalone17	Abalone20	Abalone19
RF	0.67	0.02	0.20	0.04	0.00
DT	**0.71**	0.00	0.00	0.00	0.00
DT_O	0.67	0.06	0.35	0.02	0.02
DT_U	0.69	0.08	0.33	0.18	0.00
DT_{OU}	0.62	0.08	0.31	0.15	**0.04**
LSVM	0.62	0.09	0.29	**0.21**	**0.04**
RBFSVM	0.63	**0.16**	0.17	0.00	0.00
ME^2	0.62	**0.16**	**0.37**	**0.21**	**0.04**

Performance is evaluated with respect to the F-Measure

Datasets	Abalone	Wine	Abalone17	Abalone20	Abalone19
Rate of pos. examples	10.7%	3.3%	2.5%	1.4%	0.76%

Comparaison w.r.t. a decreasing nb. of positives





Conclusion

- Capture non linearity via local models
- ME^2 is theoretically founded (uniform stability)
- Models can be learned in parallel
- Promising results in unbalanced scenarios

Future Work

- Continue the work on the decision rule
- Derive some theoretical results using the ellipsoids and Nearest Neighbor Algorithm
- Using the Global decision to improve the local decision
- Apply it on real data sets

Thank you for your attention!