Learning Maximum Excluding Ellipsoids in Unbalanced Scenarios with Theoretical Guarantees

G. Metzler[1,3], X. Badiche[3], B. Belkasmi[3], S. Canu[2], E. Fromont [1], A. Habrard[1] and M. Sebban[1]

1. Univ. Lyon, Univ. St-Etienne F-42000, UMR CNRS 5516, Laboratoire Hubert-Curien

2. LITIS EA 4108, Univ. Rouen 76800 St-Etienne du Rouvray

3. BLITZ BUSINESS SCEB, 38090 Villefontaine

Abstract

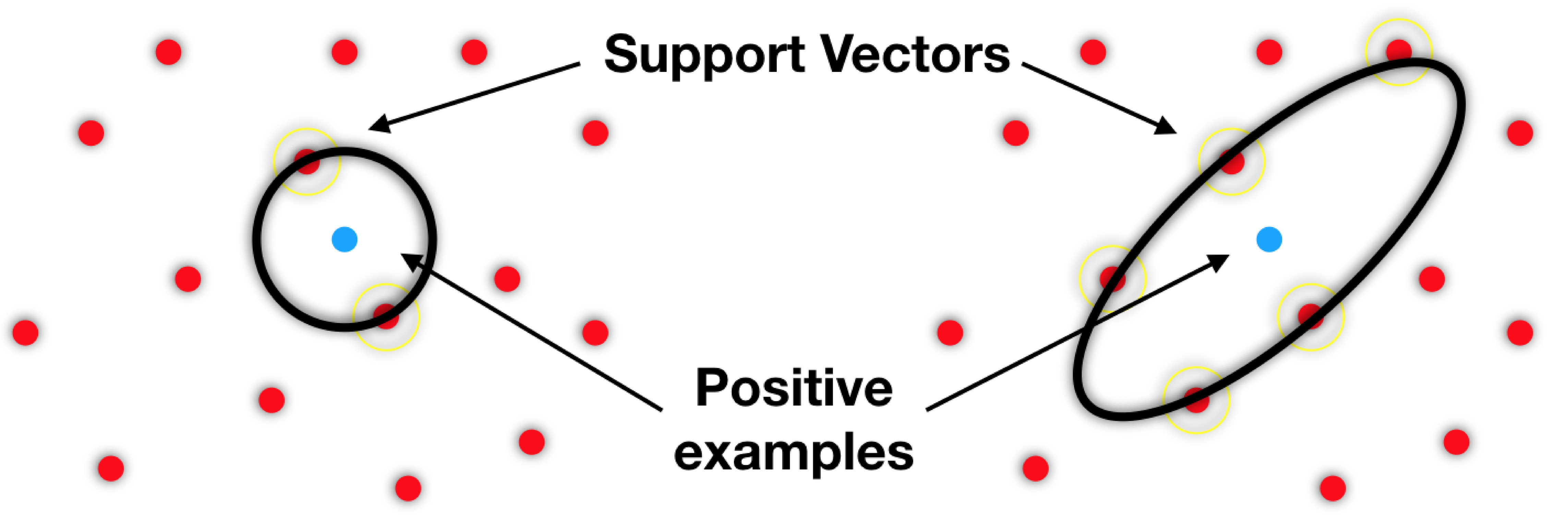
We propose a theoretically-founded method to learn maximum excluding balls in the context of unbalanced binary classification. The objective is to learn a set of local balls centered at the minority class examples which exclude the examples of the majority class. Our contribution is twofold:

1) we address this problem from a metric learning point of view [2],

2) we derive generalization guarantees on the radius of the sphere and the learned metric using the uniform stability framework [3]

Our experimental evaluation on classic benchmarks shows the effectiveness of our approach.

Illustration



From excluding ball to learned excluding ellipsoid
$\Rightarrow$ The Maximum Excluding Ellipsoids (ME^2) algorithm

Method

We learn a PSD matrix $\mathbf{M}$ of a Mahalanobis-distance defined as follows:

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

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

where,

- B is the bound of the radius,
- μ controls the size of the ellipsoid,
- λ controls the distortion w.r.t. an Euclidean ball.

Dual Form

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

The Dual Problem gives an explicit expression of both the **Radius** and the **Metric**:

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

$$\boldsymbol{M} = I + \frac{1}{2\lambda}\sum_{i=1}^n \alpha_i (x_i - c)(x_i - c)^T.$$

This expression directly shows that $\mathbf{M}$ is **PSD**.

Theoretical Bound

Definition: *A learning algorithm has a uniform stability in $\frac{\gamma}{n}$ w.r.t. a loss function ℓ and a parameter set θ, with γ a positive constant if:*

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

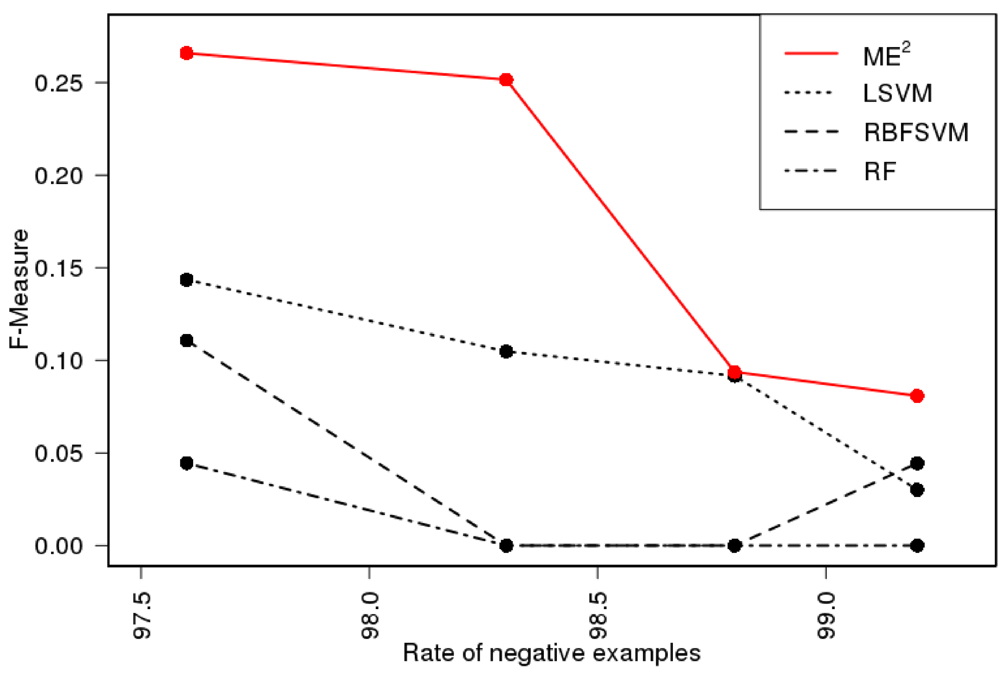
Theorem: *Let $\delta > 0$ and $n > 1$. There exists a constant $\kappa > 0$, such that with probability at least $1-\delta$ over the random draw over S, we have for any $(\mathbf{M}, R)$ solution of Problem (1):*

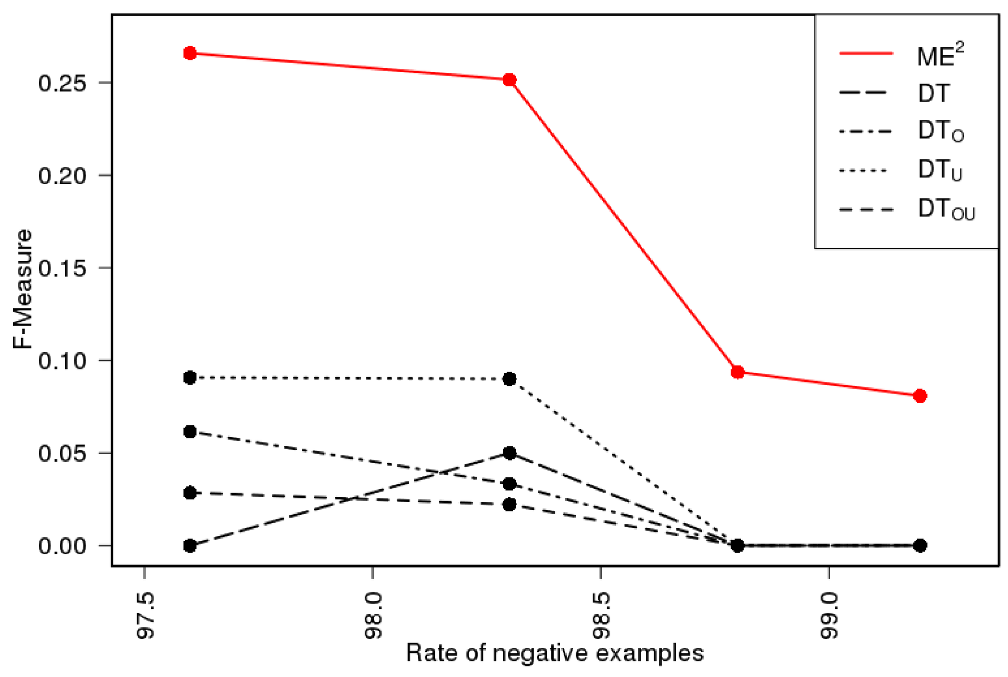
$$\begin{aligned} L(\mathbf{M}, R) \leq \hat{L}_S(\mathbf{M}, R) &+ \frac{4\max(1, 4B^2)}{n\kappa\min(\mu,\lambda)} \\ &+ \left(\frac{8\max(1,4B^2)}{\kappa\min(\mu,\lambda)} + B\right)\sqrt{\frac{\ln 1/\delta}{2n}} \\ &+ \left(4B^2\sqrt{\frac{\mu B^2}{\lambda} + d}\right)\sqrt{\frac{\ln 1/\delta}{2n}}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

Results

Data / Algo.	Abalone 10.7%	Abalone17 2.5%	Yeast6 2.4%	Abalone20 1.4%	Abalone19 0.76%
RF (10 trees)	0.67	0.20	0.04	0.00	0.00
DT (simple version)	**0.71**	0.00	0.00	0.00	0.00
DT_O (oversampling)	0.67	0.35	0.09	0.018	0.02
DT_U (undersampling)	0.69	0.33	0.09	0.18	0.00
DT_{OU} (both)	0.62	0.31	0.17	0.15	0.04
LSVM	0.62	0.29	0.15	**0.21**	0.04
RBFSVM	0.63	0.17	0.09	0.00	0.00
$\boldsymbol{ME^2}$	0.62	**0.37**	**0.26**	**0.21**	**0.04**

Comparison in terms of *F-Measure* with some state of the art algorithms.
Datasets are ordered w.r.t. to an increasing imbalance ratio. The same global weight is given to both classes to learn the LSVM and RBF-SVM.





Effectiveness of ME^2 when the disequilibrium increases. We can note that some of the state of the art methods lead to a null F-Measure.

Conclusion

The ME^2 algorithm presents the following advantages:

- Captures non linearity via local linear models
- Theoretically founded (uniform stability)
- Models can be learned in parallel
- Promising results in unbalanced scenarios

Perspective: study the link between the stability of the ellipsoids and the generalization error of a Nearest Neighbor algorithm.

References

References

[1] D. M. J. Tax and R. P. W. Duin, Support vector data description *Machine Learning Journal* (2013), vol.5, 287-364.

[2] A. Bellet, A. Habrard and M. Sebban, A survey on metric learning for feature vectors and structured data, *arXiv*, (2013)

[3] O. Bousquet and A. Elisseeff, Stability and generalization, *Journal of Machine Learning Research*, (2002) 499-526.