Poster CAP

A Corrected Nearest Neighbor Algorithm Maximizing the F-Measure from Imbalanced Data

R. Viola[1,2], R. Emonet[1], A. Habrard [1], G. Metzler[1], S. Riou[2] and M. Sebban[1]

1. Univ. Lyon, UJM-Saint-Etienne, Laboratoire Hubert Curien UMR CNRS 5516, F-42023, Saint-Etienne, France
2. DGFiP, Ministère de l'Economie et des Finances, France

Context and Notations

About the DGFiP

- Part of the French Ministry Of Economy and Finances.
- Works on the tax return frauds detection.
- 50,000 inspections per year, 3 million companies.
- A fraud ratio between 30% and 0.05%.

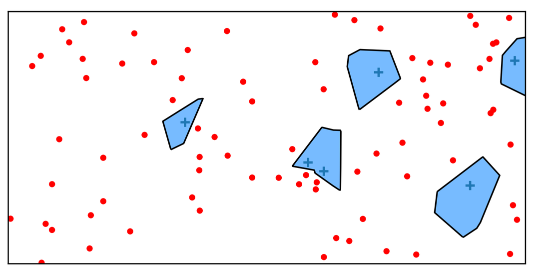
→ **Imbalanced setting**

- Use of k−NN for decision support.
- Missing fraudsters can be more expensive than inspecting non fraudsters.
- Fraudsters try to mimic non fraudsters.

→ **Moving the decision boundary**

Notations

- Use a k-NN algorithm.
- d : a distance.
- γ : a positive parameter.
- (FP, FN): False Positives and False Negatives.

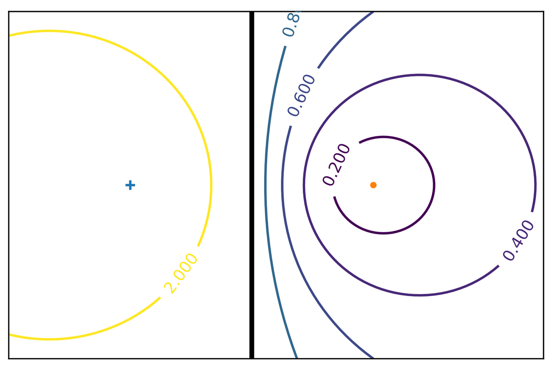


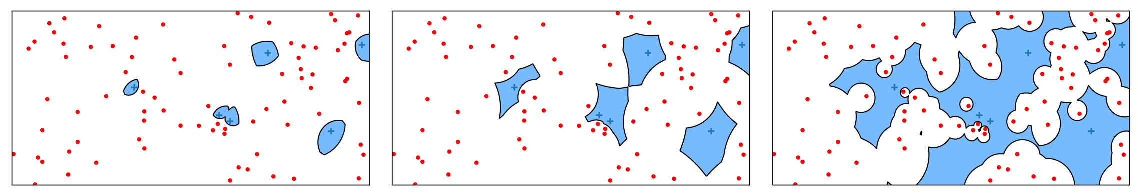
Voronoi regions

Our γk−Nearest Neighbors Algorithm

Modifying the distance to positive examples:

$$d_\gamma(\boldsymbol{x}, \boldsymbol{x}_i) = \begin{cases} d(\boldsymbol{x}, \boldsymbol{x}_i) & \text{if } \boldsymbol{x}_i \in S_-, \\ \gamma \cdot d(\boldsymbol{x}, \boldsymbol{x}_i) & \text{if } \boldsymbol{x}_i \in S_+. \end{cases}$$





$\gamma = 1.6$ $\gamma = 0.75$ $\gamma = 0.4$

→ **How to choose the best value of γ ?**

- Too big $\implies$ small Voronoi regions $\implies$ low probability to recover frauds.
- Too small $\implies$ big Voronoi regions $\implies$ high probability of false positives.

Proposition: Let ε be the distance of a query $\boldsymbol{z}$ to its nearest neighbor $\boldsymbol{x}'$. Suppose that $\gamma \leq 1$ and $(\boldsymbol{z}, \boldsymbol{x}') \in S_+ \times S_+$, then:

$$FN_\gamma(\boldsymbol{z}) = \left(1 - P(\boldsymbol{x}' \in \mathcal{S}_{\frac{\epsilon}{\gamma}}(\boldsymbol{z}))\right)^{m+} \leq (1 - P(\boldsymbol{x}' \in \mathcal{S}_\epsilon(\boldsymbol{z})))^{m+} = FN(\boldsymbol{z}).$$

If $\gamma \geq 1$ and $(\boldsymbol{z}, \boldsymbol{x}') \in S_- \times S_-$, then:

$$FP_\gamma(\boldsymbol{z}) = (1 - P(\boldsymbol{x}' \in \mathcal{S}_{\gamma\epsilon}(\boldsymbol{z})))^{m-} \leq (1 - P(\boldsymbol{x}' \in \mathcal{S}_\epsilon(\boldsymbol{z})))^{m-} = FP(\boldsymbol{z}).$$

Decreasing FN while keeping "reasonable" rate $FP \implies \boldsymbol{\gamma < 1}$.

Algorithm

Algorithm 1: Classification of a new example with γk−NN

Input : a query x to be classified, a set of labeled samples $S = S_+ \cup S_-$, a number of neighbors k, a positive real value γ, a distance function d

Output: the predicted label of x

$\mathcal{NN}^-, \mathcal{D}^- \leftarrow nn(k, \boldsymbol{x}, S_-)$ // nearest negative neighbors with their distances

$\mathcal{NN}^+, \mathcal{D}^+ \leftarrow nn(k, \boldsymbol{x}, S_+)$ // nearest positive neighbors with their distances

$\mathcal{D}^+ \leftarrow \gamma \cdot \mathcal{D}^+$

$\mathcal{NN}_\gamma \leftarrow firstK\left(k, sortedMerge((\mathcal{NN}^-, \mathcal{D}^-), (\mathcal{NN}^+, \mathcal{D}^+))\right)$

$y \leftarrow +$ if $|\mathcal{NN}_\gamma \cap \mathcal{NN}^+| \geq \frac{k}{2}$ else $-$ // majority vote based on $\mathcal{NN}_\gamma$

return y

Experimental Results with $k = 3$

DATASETS	3−NN	DUPk−NN	Wk−NN	CWk−NN	LMNN	γk−NN
BALANCE	0.954(0.017)	0.954(0.017)	0.957(0.017)	0.961(0.010)	**0.963**(0.012)	0.954(0.029)
AUTOMPG	0.808(0.077)	0.826(0.033)	0.810(0.076)	0.815(0.053)	0.827(0.054)	**0.831**(0.025)
IONOSPHERE	0.752(0.053)	0.859(0.021)	0.756(0.060)	0.799(0.036)	0.890(0.039)	**0.925**(0.017)
PIMA	0.500(0.056)	0.539(0.033)	0.479(0.044)	0.515(0.037)	0.499(0.070)	**0.560**(0.024)
WINE	0.881(0.072)	0.852(0.057)	0.881(0.072)	0.876(0.080)	**0.950**(0.036)	0.856(0.086)
GLASS	0.727(0.049)	0.733(0.061)	0.736(0.052)	0.717(0.055)	0.725(0.048)	**0.746**(0.046)
GERMAN	0.330(0.030)	0.449(0.037)	0.326(0.030)	0.344(0.029)	0.323(0.054)	**0.464**(0.029)
VEHICLE	0.891(0.044)	0.867(0.027)	0.891(0.044)	0.881(0.021)	**0.958**(0.020)	0.880(0.049)
HAYES	0.036(0.081)	0.183(0.130)	0.050(0.112)	0.221(0.133)	0.036(0.081)	**0.593**(0.072)
SEGMENTATION	0.859(0.028)	0.862(0.018)	0.877(0.028)	0.851(0.022)	**0.885**(0.034)	0.848(0.025)
ABALONE8	0.243(0.037)	0.318(0.013)	0.241(0.034)	0.330(0.015)	0.246(0.065)	**0.349**(0.018)
YEAST3	0.634(0.066)	0.670(0.034)	0.634(0.066)	**0.699**(0.015)	0.667(0.055)	0.687(0.033)
PAGEBLOCKS	0.842(0.020)	0.850(0.024)	0.849(0.019)	0.847(0.029)	**0.856**(0.032)	0.844(0.023)
SATIMAGE	0.454(0.039)	0.457(0.027)	0.454(0.039)	0.457(0.023)	**0.487**(0.026)	0.430(0.008)
LIBRAS	**0.806**(0.076)	0.788(0.187)	**0.806**(0.076)	0.789(0.097)	0.770(0.027)	0.768(0.106)
WINE4	0.031(0.069)	**0.090**(0.086)	0.031(0.069)	0.019(0.042)	0.000(0.000)	**0.090**(0.036)
YEAST6	0.503(0.302)	0.449(0.112)	0.502(0.297)	0.338(0.071)	0.505(0.231)	**0.553**(0.215)
ABALONE17	0.057(0.078)	**0.172**(0.086)	0.057(0.078)	0.096(0.059)	0.000(0.000)	0.100(0.038)
ABALONE20	0.000(0.000)	0.000(0.000)	0.000(0.000)	**0.067**(0.038)	0.057(0.128)	0.052(0.047)
MEAN	0.543(0.063)	0.575(0.053)	0.544(0.064)	0.559(0.046)	0.560(0.053)	**0.607**(0.049)

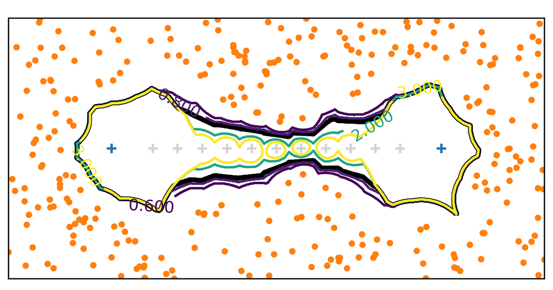
- k−NN: standard version
- DUPk−NN: with duplicated positives
- Wk−NN: weights w.r.t. distance
- CWk−NN: weights w.r.t. class
- LMNN: metric learning [1]
- γk−NN: our approach

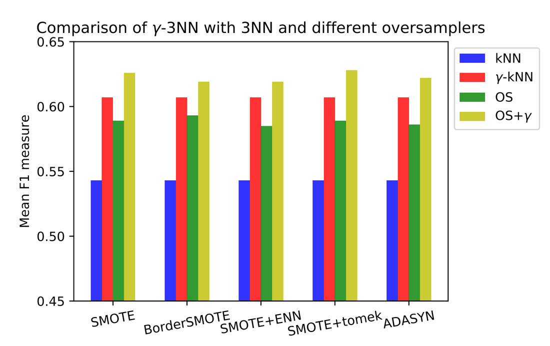
Combining γk−NN with Sampling

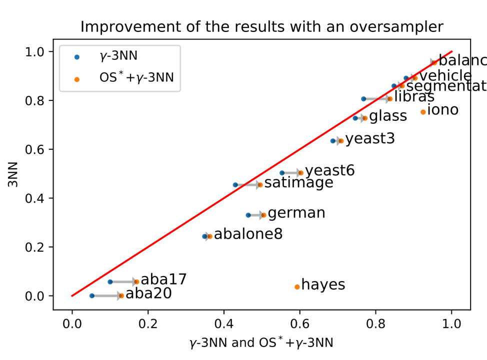
Using sampling strategies to create synthetic positives:

- Modify the distribution
- Not controllable : similar to add noise

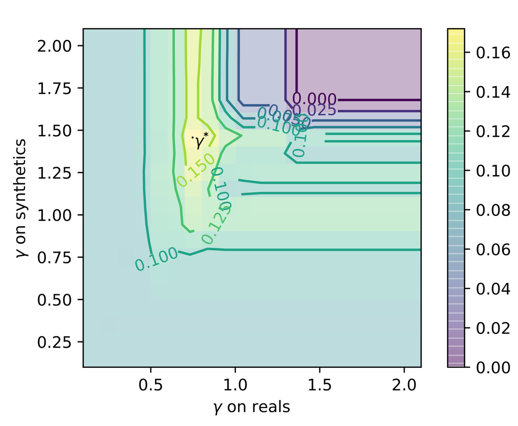
→ Use two γ: one for real and one for synthetic.







- Coupling to sampling strategy improves the results.



γ values are as expected.

- $\gamma < 1$ for real positives
- $\gamma > 1$ for synthetic ones (may be considered as noise)

Experimental Results (DGFiP)

DATASETS	3−NN	γk−NN	SMOTE	SMOTE+γk−NN
DGFIP19 2	0,454(0,007)	0,528(0,005)	0,505(0,010)	**0,529**(0,003)
DGFIP9 2	0,173(0,074)	0,396(0,018)	0,340(0,033)	**0,419**(0,029)
DGFIP4 2	0,164(0,155)	0,373(0,018)	0,368(0,057)	**0,377**(0,018)
DGFIP8 1	0,100(0,045)	**0,299**(0,010)	0,278(0,043)	**0,299**(0,011)
DGFIP8 2	0,140(0,078)	0,292(0,028)	**0,313**(0,048)	0,312(0,021)
DGFIP9 1	0,088(0,090)	0,258(0,036)	0,270(0,079)	**0,288**(0,026)
DGFIP4 1	0,073(0,101)	0,231(0,139)	0,199(0,129)	**0,278**(0,067)
DGFIP16 1	0,049(0,074)	0,166(0,065)	0,180(0,061)	**0,191**(0,081)
DGFIP16 2	0,210(0,102)	0,202(0,056)	0,220(0,043)	**0,229**(0,026)
DGFIP20 3	0,142(0,015)	0,210(0,019)	0,199(0,015)	**0,212**(0,019)
DGFIP5 3	0,030(0,012)	0,105(0,008)	**0,110**(0,109)	0,107(0,010)
MEAN	0,148(0,068)	0,278(0,037)	0,271(0,057)	**0,295**(0,028)

Impact of combining our approach, γk−NN, with a SMOTE sampling strategy [2] on the DGFiP datasets.

Perspectives

- Making our γ non stationary, i.e. having a γ which depends on the region in the feature space.
- Generalizing our algorithm using a metric learning approach.
- Derive generalization guarantees.

References

[1] Weinberger, K. Q., & Saul, L. K. (2009). *Distance metric learning for large margin nearest neighbor classification.*

[2] Chawla, N. V., Bowyer, K. W., Hall, L. O., & Kegelmeyer, W. P. (2002). *SMOTE: synthetic minority over-sampling technique.*