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

R. Viola, R. Emonet, A. Habrard, G. Metzler, S. Riou, M. Sebban

July 4, 2019

Context

Context

Fraud detection for the French Ministry of Economy and Finance (DGFiP).

Context

Context

Fraud detection for the French Ministry of Economy and Finance (DGFiP).

Imbalanced datasets

A fraud rate between 0.05% to 30%.

Context

Context

Fraud detection for the French Ministry of Economy and Finance (DGFiP).

Imbalanced datasets

A fraud rate between 0.05% to 30%.

Nearest neighbors

The nearest neighbor algorithm is used by the DGFiP for decision support.

Observations

Limits

The small number of positives leads to little areas of positive prediction.

Observations

Limits

The small number of positives leads to little areas of positive prediction.

Needs

Reduce the number of undetected fraudsters without increasing the number of false alarms too much.

Observations

Limits

The small number of positives leads to little areas of positive prediction.

Needs

Reduce the number of undetected fraudsters without increasing the number of false alarms too much.

Ideas

Increase the area of influence of the positives.

Observations

Limits

The small number of positives leads to little areas of positive prediction.

Needs

Reduce the number of undetected fraudsters without increasing the number of false alarms too much.

Ideas

Increase the area of influence of the positives.

schemas/infl-thr-100.pdf	schemas/infl-thr-075.pdf	schemas/infl-thr-040.

Proposed method

$\gamma k-$NN

Scale the distance between a query point and positive training examples by a factor.

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

Algorithm:Classification of a new example

Input : a query $\mathbf{x}$ 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 $\mathbf{x}$

$\mathcal{NN}^-, \mathcal{D}^- \leftarrow nn(k, \mathbf{x}, S_-)$ // nearest negative neighbors with their distances
$\mathcal{NN}^+, \mathcal{D}^+ \leftarrow nn(k, \mathbf{x}, S_+)$ // 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 $\left|\mathcal{NN}_\gamma \cap \mathcal{NN}^+\right| \geq \frac{k}{2}$ else $-$ // majority vote based on $\mathcal{NN}_\gamma$
return y

Theoretical analysis

Proposition on False Negative and False Positive probability

$FN_\gamma(\mathbf{z})$ the probability for a positive example $\mathbf{z}$ to be a false negative using our algorithm. If $\gamma \leq 1$,

$$FN_\gamma(\mathbf{z}) \quad \leq \quad FN(\mathbf{z})$$

Theoretical analysis

Proposition on False Negative and False Positive probability

$FN_\gamma(\mathbf{z})$ the probability for a positive example $\mathbf{z}$ to be a false negative using our algorithm. If $\gamma \leq 1$,

$$FN_\gamma(\mathbf{z}) \quad \leq \quad FN(\mathbf{z})$$

Proposition on False Positive probability

$FP_\gamma(\mathbf{z})$ the probability for a negative example $\mathbf{z}$ to be a false positive using our algorithm. If $\gamma \geq 1$,

$$FP_\gamma(\mathbf{z}) \quad \leq \quad FP(\mathbf{z})$$

Theoretical analysis(Ctd.)

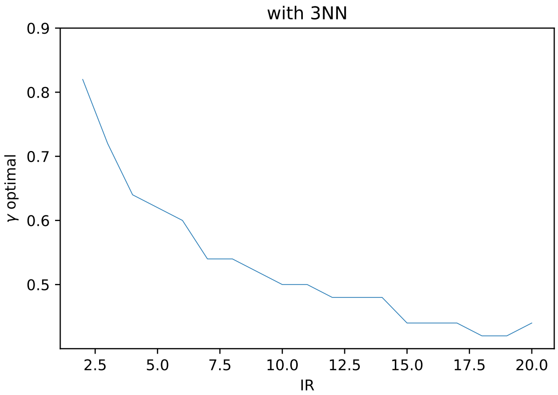


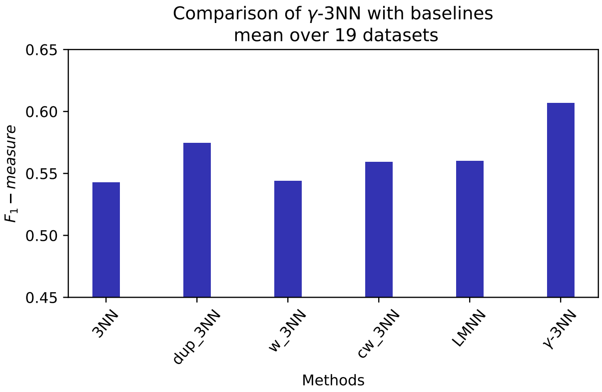
Figure: Evolution of the optimal γ value with respect to the IR for $k = 3$.

Experiments and results

- Comparison with 5 baselines.
- Tests on 19 public datasets and 11 private datasets.
- 10 CV of the value of γ and Mean of 5 experiments.

Experiments and results

- Comparison with 5 baselines.
- Tests on 19 public datasets and 11 private datasets.
- 10 CV of the value of γ and Mean of 5 experiments.

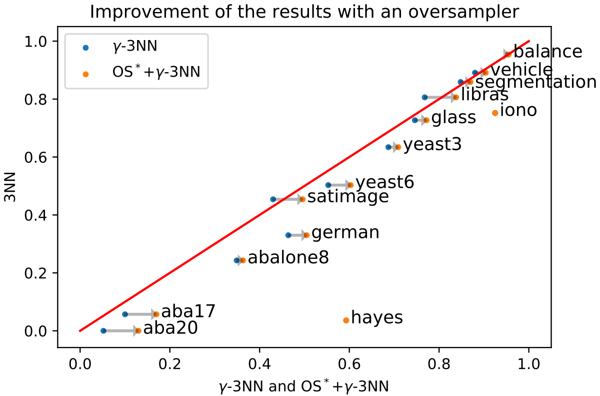


Experiments and results(Ctd.)

- As a complement to an oversampler.
- Selection of the best ratio between the actual one and 1.
- 2 separated values of γ (real and synthetic positives).

Experiments and results(Ctd.)

- As a complement to an oversampler.
- Selection of the best ratio between the actual one and 1.
- 2 separated values of γ (real and synthetic positives).



Experiments and results(Ctd.)

Table: Results for $3-$NN on the DGFiP datasets.

DATASETS	3−NN	$\gamma k-$NN	SMOTE	SMOTE+$\gamma k-$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)

Future Work

Lines of research

- 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.

Questions ?