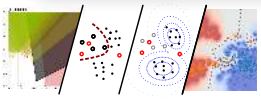
Learning from Imbalanced Data and Anomaly Detection

Rémi Emonet

Université Jean-Monnet,
Laboratoire Hubert Curien,
Saint-Étienne

Talk at SLEIGHT Science Event#6, 2021-07-06

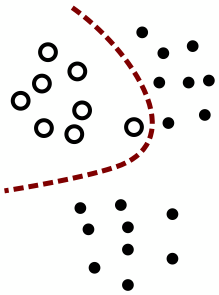


Overview

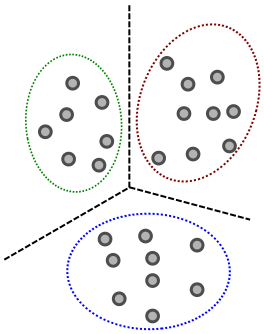
- Introduction
- Anomaly and (rare) event detection
  - Problem, notations and performance measures
- Imbalanced classification problems
  - General approaches
  - Correcting k-NN: γ-***NN*** and ***MLFP***
  - Learning Maximum Excluding Ellipsoids
  - Focusing on the F-Measure optimization
- Unsupervised anomaly detection
  - Simple motivating approach
  - A variety of ML methods
  - Probabilistic models 101
  - Case study: temporal motif mining
  - More models (VAE, GAN, tensor networks, ...)
- Closing remarks

Supervised / Unsupervised *Machine Learning*

Supervised
(known labels)



Unsupervised
(only inputs)



Supervised vs Unsupervised Learning

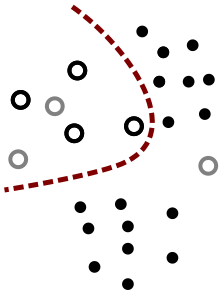
- Supervised
  - Given example inputs (X) and corresponding outputs (y)
  - Learn the function "input $\rightarrow$ output" ($y = f(X)$)
    - classification (categorical output)
    - regression (continuous output)
  - Methods:
    k-NN, SVM, SVR, Random Forests, Least Squares fit,
    Neural Networks, Gaussian Processes, Boosting, ...
- Unsupervised
  - Given a set of data points (X)
  - Model/structure/understand this dataset
    - clustering, density estimation
    - source separation
    - pattern and sequence mining
    - rare events / anomaly detection
  - Methods:
    PCA, k-means, DBSCAN, OneClass-SVM, Isolation Forests,
    Prob. mod. (GMM, HMM, GPLVM, ...), Autoencoders, GANs, KDE ...

Focus on k-NN (ML reminder)

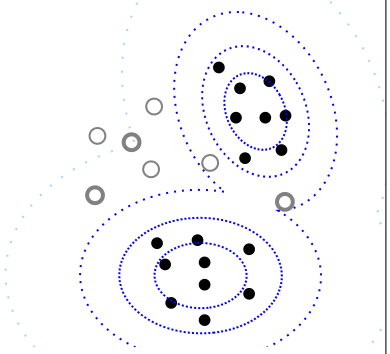
Focus on PCA (ML reminder)

Anomaly Detection *supervised or unsupervised*

Supervised
(some anomalies)



Unsupervised
(only normal points)



Disclaimers

The bibliography is not exhaustive, see references in the mentioned papers

The presentation is not exhaustive, it is a tour of some methods

The slides contain some (scary) details that I won't cover

Overview

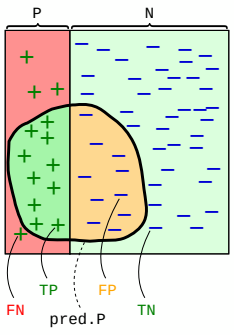
- Introduction
- Anomaly and (rare) event detection
  - Problem, notations and performance measures
- Imbalanced classification problems
  - General approaches
  - Correcting k-NN: γ-***NN*** and ***MLFP***
  - Learning Maximum Excluding Ellipsoids
  - Focusing on the F-Measure optimization
- Unsupervised anomaly detection
  - Simple motivating approach
  - A variety of ML methods
  - Probabilistic models 101
  - Case study: temporal motif mining
  - More models (VAE, GAN, tensor networks, ...)
- Closing remarks

Event Detection Problems

- Anomaly detection
  - unsafe situations in videos
  - defect detection in images
  - abnormal heart beat detection in ECG
- Fraud detection
  - fraudulent checks
  - credit card fraud (physical, online)
  - financial fraud (French DGFIP)

Imbalanced Classification Problems

- Binary classification
  - $+$ positive class: minority class, anomaly, rare event, ...
  - $-$ negative class: majority class, normality, typical event, ...
- Confusion matrix (of a model vs a ground truth)
  - TP: true positive ... $TP = P - FN$
  - FP: false positive
  - TN: true negative ... $TN = N - FP$
  - FN: false negative
- Some measures
  - Precision: $prec = \frac{TP}{TP + FP}$
  - Recall: $rec = \frac{TP}{P} = \frac{TP}{TP + FN}$
  - F_β-measure: $F_\beta = (1 + \beta^2)\frac{prec \cdot rec}{\beta^2 \cdot prec + rec}$

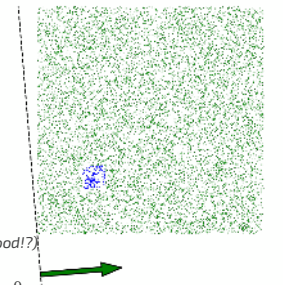


F-measure vs Accuracy ?

F_β-measure: $F_\beta = (1 + \beta^2)\frac{prec \cdot rec}{\beta^2 \cdot prec + rec}$

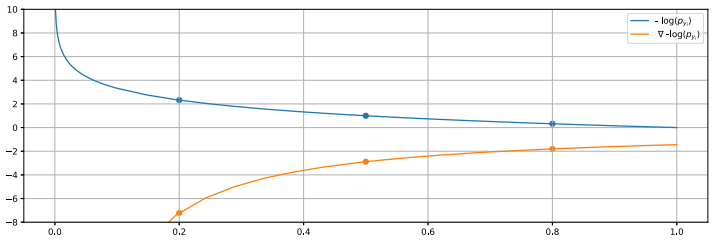
$accuracy = \frac{TP + TN}{P + N}$

- Accuracy inadequacy
  - with $P = 100, N = 10000$
    with a classifier that always says "*normal*":
    $TP = 0, TN = N, FP = 0, FN = P$
    - $accuracy = \frac{0 + N}{P + N} = \frac{10000}{10100} = 99\%$ *(very good!?)*
    - $F_\beta = (1 + \beta^2)\frac{prec \cdot rec}{\beta^2 \cdot prec + rec} = (1 + \beta^2)\frac{1 \cdot 0}{\beta^2 \cdot 1 + 0} = 0$ *(very bad!... but that's right)*



- F_β-measure challenges
  - discrete (like the accuracy)
  - non-convex (even with continuous surrogates)
  - **non-separable**, i.e. $F_\beta \neq \sum_{(x_i, y_i) \in S} \ldots$
    ⇒ The loss for one point depends on the others
    ⇒ Impossible to optimize directly
    ⇒ Impossible to optimize on a subset (minibatch)

Ok, but... I'm doing gradient descent, so...



- Gradient: $0.2 \Rightarrow -7.21$, $0.5 \Rightarrow -2.89$, $0.8 \Rightarrow -1.80$, $1 \Rightarrow -1.44$
- Example, gradient intensity is the same for:
  - $10+$ wrongly classified with an output proba. of 0.2
  - $40-$ correctly classified with an output proba 0.8
  - i.e., lazily predicting systematically 0.2 (for $+$) yields a "stable" solution with $10+$ vs $40-$

Ok, but... my deep model does 100%...

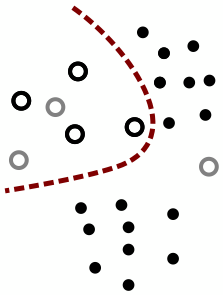
- ... the 100% accuracy is on the train set
- ... I cannot tell you if it will generalize well

Overview

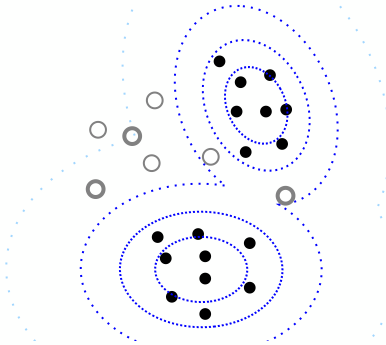
- Introduction
- Anomaly and (rare) event detection
  - Problem, notations and performance measures
- Imbalanced classification problems
  - General approaches
  - Correcting k-NN: γ-***NN*** and ***MLFP***
  - Learning Maximum Excluding Ellipsoids
  - Focusing on the F-Measure optimization
- Unsupervised anomaly detection
  - Simple motivating approach
  - A variety of ML methods
  - Probabilistic models 101
  - Case study: temporal motif mining
  - More models (VAE, GAN, tensor networks, ...)
- Closing remarks

Reminder: Class Imbalance or Novelty?

Supervised
(some anomalies)



Unsupervised
(only normal points)



Overview

- Introduction
- Anomaly and (rare) event detection
  - Problem, notations and performance measures
- Imbalanced classification problems
  - General approaches
  - Correcting k-NN: γ-***NN*** and ***MLFP***
  - Learning Maximum Excluding Ellipsoids
  - Focusing on the F-Measure optimization
- Unsupervised anomaly detection
  - Simple motivating approach
  - A variety of ML methods
  - Probabilistic models 101
  - Case study: temporal motif mining
  - More models (VAE, GAN, tensor networks, ...)
- Closing remarks

Counteracting Imbalance

... when we have a strong majority of $-$ (negative examples).

- Undersampling the majority class $-$
  - throw away some points
- Oversampling class $+$
  - duplicate some points
- Generating fake $+$
  - generate between existing points (SMOTE, ...)
  - generate "around" existing points
- Using a weighted-classifiers learner
  - optimize balanced accuracy
  - challenge: the optimal weight is unknown

Overview

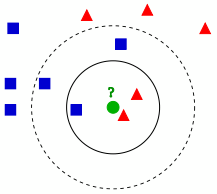
- Introduction
- Anomaly and (rare) event detection
  - Problem, notations and performance measures
- Imbalanced classification problems
  - General approaches
  - Correcting k-NN: γ-***NN*** and ***MLFP***
  - Learning Maximum Excluding Ellipsoids
  - Focusing on the F-Measure optimization
- Unsupervised anomaly detection
  - Simple motivating approach
  - A variety of ML methods
  - Probabilistic models 101
  - Case study: temporal motif mining
  - More models (VAE, GAN, tensor networks, ...)
- Closing remarks

An Adjusted Nearest Neighbor Algorithm Maximizing the F-Measure from Imbalanced Data

- **Rémi Viola**, Rémi Emonet, Amaury Habrard, **Guillaume Metzler**, Sébastien Riou, Marc Sebban
- ICTAI2019 (+ long version in IJAIT 2021)

k-NN: k Nearest Neighbor Classification

- k-NN
  - to classify a new point
  - find the closest k points (in the training section)
  - use a voting scheme to affect a class
  - efficient algorithms (K-D Tree, Ball Tree)

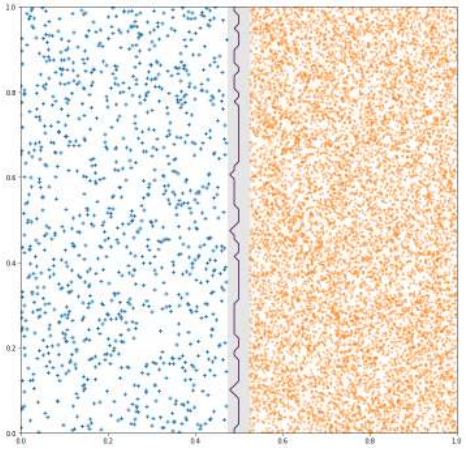


- Does k-NN still matter?
  - non-linear by design (with similarity to RBF-kernel SVM)
  - no learning, easy to patch a model (add/remove points)
  - Limits of k-NN for imbalanced data?

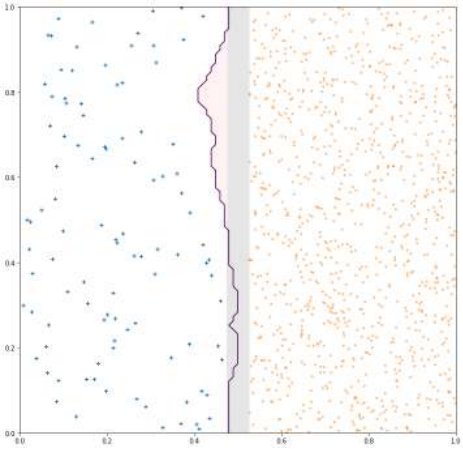
Limits of k-NN for imbalanced data?

1. k-NN behavior in uncertain areas
   - i.e., for some feature vector, the class can be $+$ or $-$
   - i.e., the "Risk of the Optimal Bayes Classifier" is non zero
   - ✔ not so bad, 1-NN respects imbalance (not k-NN)
2. k-NN behavior around boundaries
   - i.e., what happens if classes are separate but imbalanced
   - ✖ sampling effects cause problems

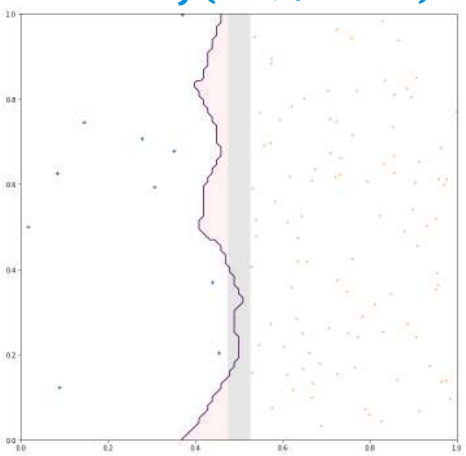
1-NN at a boundary (1000 $+$ / 10k $-$)



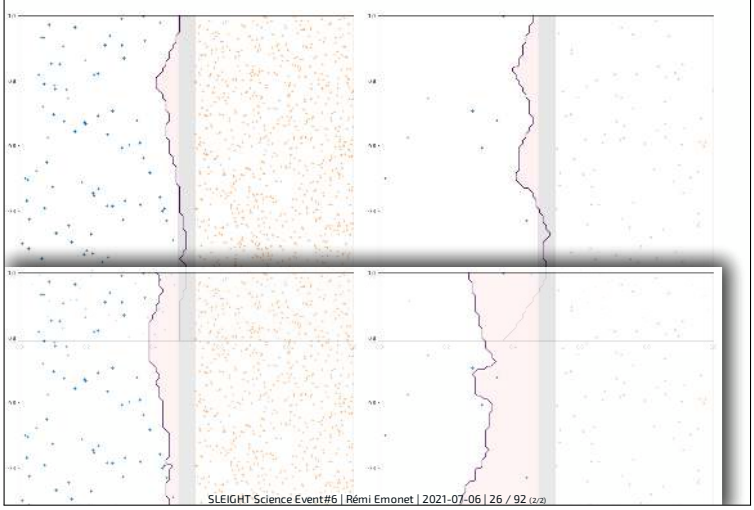
1-NN at a boundary (100 $+$ / 1000 $-$)



1-NN at a boundary (10 $+$ / 100 $-$)



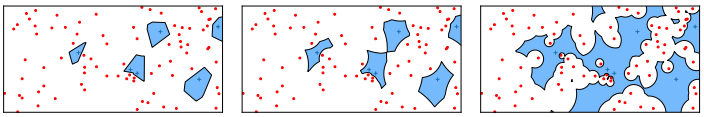
k-NN: increasing k? *(from 1 to 11) ... makes it worse*



An Adjusted Nearest Neighbor Algorithm Maximizing the F-Measure from Imbalanced Data

- **Rémi Viola**, Rémi Emonet, Amaury Habrard, **Guillaume Metzler**, Sébastien Riou, Marc Sebban
- ICTAI2019 (+ long version in IJAIT 2021)

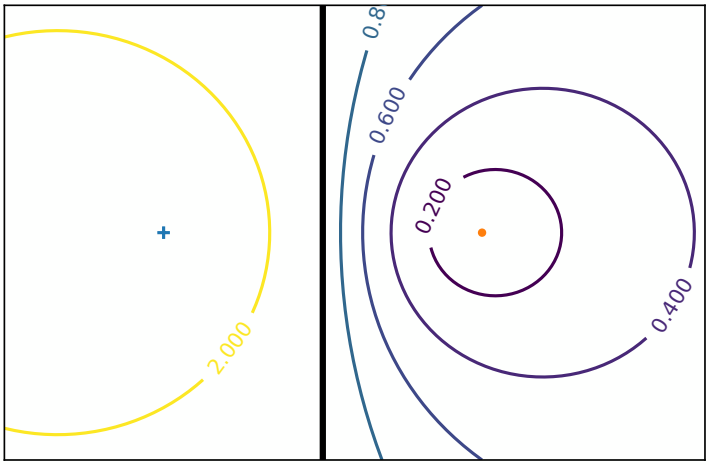
γ-NN Idea: push the decision boundary



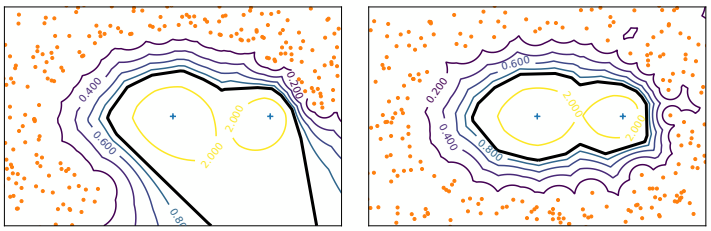
- Goal: correct for problems due to sampling with imbalance
- Genesis: GAN to generate "$+$" around existing ones
  ⇒ unstable, failing, complex
- Approach
  - artificially make $+$ closer to new points
  - how? by using a different distance for $+$ and $-$
  - the base distance to $+$ gets multiplied by a parameter γ
    (intuitively $\gamma \leq 1$ if $+$ is rare)

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

γ-NN: varying γ with two points



γ-NN: varying γ with a few $+$



- γ-NN can control
  how close to the minuses it pushes the boundary

γ-NN: Algorithm

Algorithm 1: Classification of a new example with $\gamma k-$NN

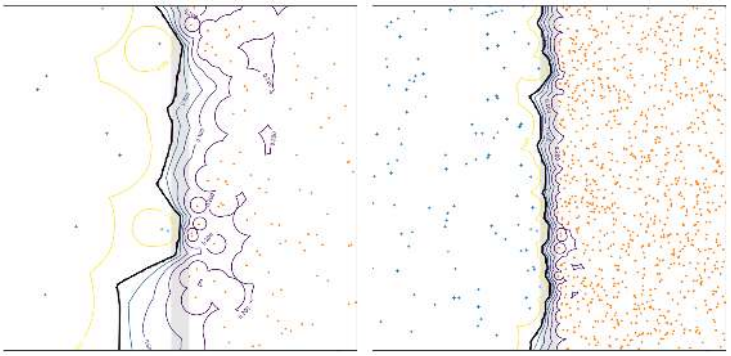
Input : a query $\mathbf{x}$ to be classified, a set of labeled samples $S = S_+ \cup S_-$, a number of neighbors k, a positive real value γ, a distance function d
Output: the predicted label of $\mathbf{x}$

$\mathcal{NN}^-, \mathcal{D}^- \leftarrow nn(k, \mathbf{x}, S_-)$ // 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 $|\mathcal{NN}_\gamma \cap \mathcal{NN}^-| \geq \frac{k}{2}$ else $-$ // majority vote based on $\mathcal{NN}_\gamma$
return y

- Trivial to implement
- Same complexity as k-NN (at most twice)
- Training
  - none, as k-NN
  - γ is selected by (cross-)validation (on the measure of interest)

γ-NN: a way to reweight distributions

- ✔ In uncertain regions (1-NN is already ok)
- ✔ At the **boundaries** (10 and 100 $+$)



Results on public datasets (F-measure)

DATASETS	$3-$NN	DUP$k-$NN	W$k-$NN	CW$k-$NN	LMNN	$\gamma k-$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)**
IONO	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 (0.136)
WINE4	0.031 (0.060)	**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 (0.067)
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)**

Results on DGFiP datasets (F-measure)

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)

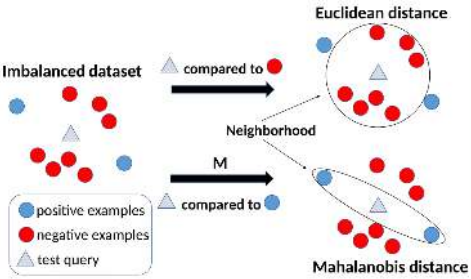
Extension With Metric Learning

- Note:
  - γ-NN learns a metric for comparing a query to a $+$
  - γ-NN kind of learn the size of a sphere around $+$
  - this is a very simple "Metric Learning"
- Extension
  - learn a full metric (a matrix M and not only a scalar γ)
  - derive a learning algorithm (not just using a validation set)

Learning from Few Positives: a Provably Accurate Metric Learning Algorithm to deal with Imbalanced Data

- **Rémi Viola**, Rémi Emonet, Amaury Habrard, **Guillaume Metzler**, Marc Sebban
- IJCAI 2020 *(International Joint Conference on Artificial Intelligence)*
  - learn a full metric (a matrix M and not only a scalar γ)
  - derive a learning algorithm (not just using a validation set)
  - derive a theoretical guarantees

MLFP: Metric Learning from Few Positive



$$\min_{\mathbf{M}\in\mathbb{S}^+} \frac{1}{m^3}\left((1-\alpha)\sum_{\substack{(\mathbf{x}_i,\mathbf{x}_j,\mathbf{x}_k)\\ y_i=y_j=1\neq y_k}} \ell_{\text{FN}}(\mathbf{M},\mathbf{z}_i,\mathbf{z}_j,\mathbf{z}_k) + \alpha \sum_{\substack{(\mathbf{x}_i,\mathbf{x}_j,\mathbf{x}_k)\\ y_i=y_j=-1\neq y_k}} \ell_{\text{FP}}(\mathbf{M},\mathbf{z}_i,\mathbf{z}_j,\mathbf{z}_k)\right) + \mu\|\mathbf{M}-\mathbf{I}\|_F^2,$$

such that $\lambda_{\max}(\mathbf{M}) \leq 1$.

where $\mathbb{S}^+$ is the set of PSD matrices, $\lambda_{\max}(\mathbf{M})$ is the largest eigenvalue of the PSD matrix $\mathbf{M}$, ℓ_{FN} and ℓ_{FP} are defined by:

$$\ell_{\text{FN}}(\mathbf{M},\mathbf{z}_i,\mathbf{z}_j,\mathbf{z}_k) = [1-c+d_{\mathbf{M}}(\mathbf{x}_i,\mathbf{x}_j)^2 - d(\mathbf{x}_i,\mathbf{x}_k)^2]_+,$$
$$\ell_{\text{FP}}(\mathbf{M},\mathbf{z}_i,\mathbf{z}_j,\mathbf{z}_k) = [1-c+d(\mathbf{x}_i,\mathbf{x}_j)^2 - d_{\mathbf{M}}(\mathbf{x}_i,\mathbf{x}_k)^2]_+,$$

Overview

- Introduction
- Anomaly and (rare) event detection
  - Problem, notations and performance measures
- Imbalanced classification problems
  - General approaches
  - Correcting k-NN: γ-***NN*** and ***MLFP***
  - Learning Maximum Excluding Ellipsoids
  - Focusing on the F-Measure optimization
- Unsupervised anomaly detection
  - Simple motivating approach
  - A variety of ML methods
  - Probabilistic models 101
  - Case study: temporal motif mining
  - More models (VAE, GAN, tensor networks, ...)
- Closing remarks

Learning maximum excluding ellipsoids from imbalanced data with theoretical guarantees

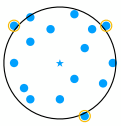
- **Guillaume Metzler**, Xavier Badiche, Brahim Belkasmi, Elisa Fromont, Amaury Habrard, Marc Sebban
- PRL2018 (Pattern Recognition Letters)
- (slides borrowed from Guillaume Metzler Ph.D. defense)

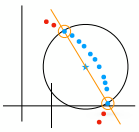
ME^2: Learning Risky Areas

Hypothesis

Frauds are close to each other, they form small groups in the feature space

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





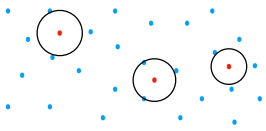
$$\min_{\mathbf{c},R,\xi} \quad R^2 + \frac{\mu}{m}\sum_{i=1}^m \xi_i,$$
$$\text{s.t.} \quad \|\mathbf{x}_i - \mathbf{c}\|_2^2 \leq R^2 + \xi_i,\ \forall i,$$
$$\xi_i \geq 0\ \forall i.$$

$$\min_{\mathbf{c},\rho,\xi} \quad \frac{1}{2}\|\mathbf{c}\|_2^2 + \frac{1}{\nu m}\sum_{i=1}^m \xi_i - \rho - \frac{1}{2}\|\mathbf{x}_i\|_2^2,$$
$$\text{s.t.} \quad \mathbf{c}^T\mathbf{x}_i \geq \rho + \frac{1}{2}\|\mathbf{x}_i\ \forall i,$$
$$\xi_i \geq 0\ \forall i.$$

Being in the ball $\iff$ being above the hyperplane

ME^2: Learning Risky Areas

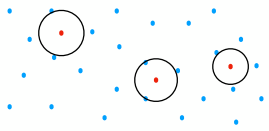
From MIB to ME^2



- Use the idea of MIB to create MEB
- One model per positive instance
- Require few positive neighbors

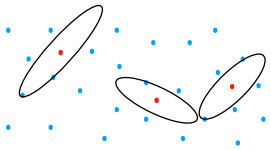
ME^2: Learning Risky Areas

From MIB to ME^2



- Use the idea of MIB to create MEB
- One model per positive instance
- Require few positive neighbors

↓ Learning a Metric



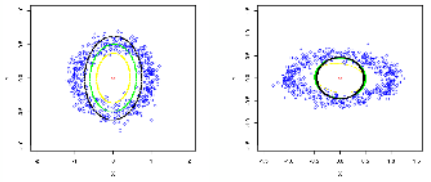
- From balls to ellipsoids
- Increase decision boundary

→ Maximum Excluding Ellipsoids

ME^2: Learning Risky Areas
Optimization problem

$$\begin{array}{ll}\min\limits_{R,\mathbf{M},\boldsymbol{\xi}} & \frac{1}{m}\sum_{i=1}^{m}\xi_i + \mu(B-R)^2 + \lambda\|\mathbf{M}-\mathbf{I}\|_{\mathcal{F}^2}, \\ s.t. & \|\mathbf{x}_i-\mathbf{c}\|^2_{\mathbf{M}} \geq R-\xi_i, \quad \forall i=1,\ldots,m, \\ & \xi_i \geq 0, \quad \forall i=1,\ldots,m \\ & 0 \leq R \leq B,\end{array}$$

error terms (in terms of distances) regularization term



Influence of μ Influence of λ

ME^2: Learning Risky Areas
Dual formulation

→ express the Lagrangian $\mathcal{L}$ including the constraints
→ expression of primal variables w.r.t. dual ones:

1. derivative of $\mathcal{L}$ w.r.t. primal variables
2. set derivatives to 0

ME^2: Learning Risky Areas
Dual formulation

→ express the Lagrangian $\mathcal{L}$ including the constraints
→ expression of primal variables w.r.t. dual ones:

1. derivative of $\mathcal{L}$ w.r.t. primal variables
2. set derivatives to 0

One of these derivatives gives:

$$\frac{\partial\mathcal{L}}{\partial\mathbf{M}} = 0 \implies \mathbf{M} = \mathbf{I} + \frac{1}{2\lambda}\sum_{i=1}^{m}\alpha_k(\mathbf{x}_k-\mathbf{c})(\mathbf{x}_k-\mathbf{c})^T.$$

→ **M is Positive Semi Definite for free**

ME^2: Learning Risky Areas
Theoretical Guarantees

Using stability framework (Bousquet and Elisseeff, 2002)

$$\mathcal{R}(\mathbf{M},R) \leq \mathcal{R}_S(\mathbf{M},R) + \mathcal{O}\left(\frac{1}{\min(\mu,\lambda)}\sqrt{\frac{\ln(1/\delta)}{2m}}\right),$$

where $\mathcal{R}_S(\mathbf{M},R) = \frac{1}{m}\sum_{i=1}^{m}[R-\|\xi-\mathbf{c}\|^2_{\mathbf{M}}]_+$.

ME^2: Learning Risky Areas
Theoretical Guarantees

Using stability framework (Bousquet and Elisseeff, 2002)

$$\mathcal{R}(\mathbf{M},R) \leq \mathcal{R}_S(\mathbf{M},R) + \mathcal{O}\left(\frac{1}{\min(\mu,\lambda)}\sqrt{\frac{\ln(1/\delta)}{2m}}\right),$$

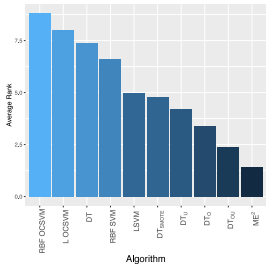
where $\mathcal{R}_S(\mathbf{M},R) = \frac{1}{m}\sum_{i=1}^{m}[R-\|\xi-\mathbf{c}\|^2_{\mathbf{M}}]_+$.

the true risk on the underlying and unknown distribution
the empirical risk over the sample S
generalization gap of the learned model: depends on the complexity of the model

ME^2: Learning Risky Areas
Experimental Results

Comparison with standards algorithms on imbalanced datasets

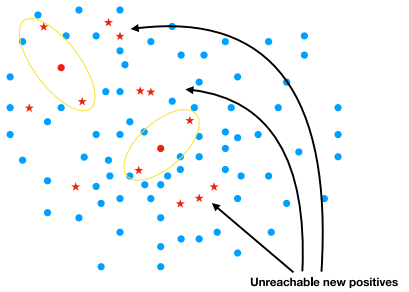
Dataset	Nb. of ex.	% Pos.
Wine	1 599	3.3
Abalone17	2 338	2.5
Yeast6	1 484	2.4
Abalone20	1 916	1.4
Blitz	15 000	1.0



Lower Rank: able to reach better performance

ME^2: Learning Risky Areas

Limitation of ME^2



Find a way to increase the influence zone of positives

Overview

- Introduction
- Anomaly and (rare) event detection
  - Problem, notations and performance measures
- Imbalanced classification problems
  - General approaches
  - Correcting k-NN: γ-***NN*** and ***MLFP***
  - Learning Maximum Excluding Ellipsoids
  - Focusing on the F-Measure optimization
- Unsupervised anomaly detection
  - Simple motivating approach
  - A variety of ML methods
  - Probabilistic models 101
  - Case study: temporal motif mining
  - More models (VAE, GAN, tensor networks, ...)
- Closing remarks

Back to the F_β-measure (or "F-score")

- Recap
  - FP *(false positive)* FN *(false negative)* $TP = P - FN$
  - $prec = \frac{TP}{TP+FP}$ and $rec = \frac{TP}{P} = \frac{TP}{TP+FN}$
  - F_β-measure: $F_\beta = (1+\beta^2)\frac{prec \cdot rec}{\beta^2 \cdot prec + rec}$
- F_β-measure challenges
  - discrete (like the accuracy)
  - non-convex (even with continuous surrogates)
  - **non-separable**, i.e. $F_\beta \neq \sum_{(x_i,y_i)\in S} \dots$

    ⇒ The loss for one point depends on the others
    ⇒ Impossible to optimize directly
    ⇒ Impossible to optimize on a subset (minibatch)

Link with Weighted classification *(intuition)*

We can rewrite

$$F_1 = 2\frac{prec \cdot rec}{prec + rec} = 2/\left(\frac{prec + rec}{prec \cdot rec}\right) = 2/\left(\frac{1}{rec} + \frac{1}{prec}\right)$$

$$F_1 = 2/\left(\frac{TP+FN}{TP} + \frac{TP+FP}{TP}\right) \qquad \text{with } P = FN$$

$$F_1 = 2/\left(\frac{2P - FN + FP}{P - FN}\right) = \frac{2(P-FN)}{2P - FN + FP} = \frac{A}{B}$$

Link with Weighted classification

Weighted classification for F_β

$$F_\beta = \frac{(1+\beta^2)\cdot(P-FN)}{1+\beta^2 P - FN + FP} = \frac{(1+\beta^2)\cdot(P-e_1)}{1+\beta^2 P - e_1 + e_2}$$

- The F_β-measure is linear fractional *(in $e = (e_1, e_2) = (FN, FP)$)*
  i.e. $F_\beta = \frac{\langle a', e\rangle + b}{\langle c, e\rangle + d} = \frac{A}{B}$
- Relation to weighted classification

  $F_\beta \geq t$ (we achieve a good, above F_β, value)
  $\Leftrightarrow A \geq t \cdot B$
  $\Leftrightarrow A - t \cdot B \geq 0$
  $\Leftrightarrow (1+\beta^2)\cdot(P-e_1) - t(1+\beta^2 P - e_1 + e_2) \geq 0$
  $\Leftrightarrow (-1-\beta^2+t)e_1 - te_2 \geq -P(1+\beta^2) + t(1+\beta^2 P)$
  $\Leftrightarrow (1+\beta^2-t)e_1 + te_2 \leq -P(1+\beta^2) + t(1+\beta^2 P)$

  ⇒ so, we can minimize the weighted problem with class weights $(1+\beta^2-t, t)$

Optimizing F-measures by cost-sensitive classification

- **Shameem P. Parambath**, Nicolas Usunier, Yves Grandvalet
- NIPS 2014 *(Advances in Neural Information Processing Systems)*
  - Main result (reformulation)
    - given a value of t (and class weights $a(t) = (1+\beta^2-t, t)$)
    - given a weighted-classification learner
    - then (thanks to some derived bound)
      the optimal $F^\star$ is not too far from the obtained F
  - They conclude that a grid search on t is sound
    i.e., *if we try enough t, we'll get arbitrary close to the optimal F_β-measure (and they experiment with 19 t values)*

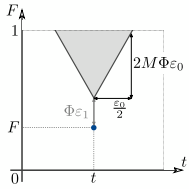
From Cost-Sensitive Classification to Tight F-measure Bounds

*Also known as **CONE***

- **Kevin Bascol**, Rémi Emonet, Élisa Fromont, Amaury Habrard, **Guillaume Metzler**, Marc Sebban
- AISTATS2019
  - revisiting Parambath2014
  - adds a geometric interpretation
  - derive **tighter** bounds on "what if I change t"
  - use the bound in an iterative algorithm

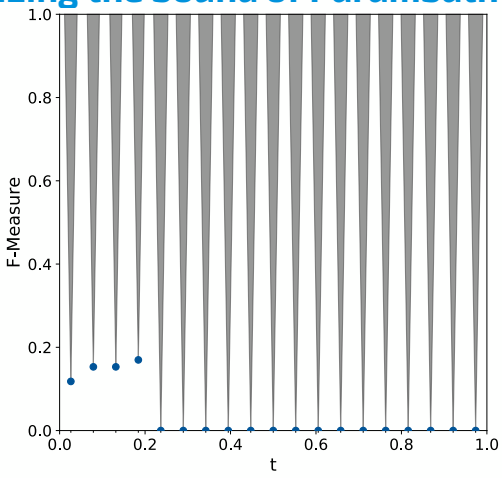
Geometric interpretation of bounds

- A reinterpretation of the bound
  - given
    - a value of t (and class weights $a(t)$), and
    - a weighted-classifier h (learned with weight $a(t)$)
    - … yielding an error profile e and a correponding $F_\beta(e)$
  - **then***, for any t' (that would yield a classifier e'),
    - *(if t' is close to t then $F(e')$ is close to $F(e)$)*
    - $F(e')$ is upper bounded by an expression linear in $|t - t'|$

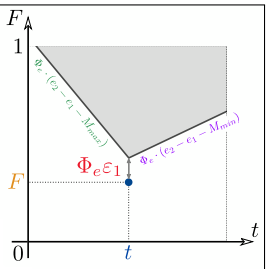


* All bounds are up to ε_1, the sub-optimality of the learned weighted classifier

Visualizing the bound of Parambath et al.



Deriving a tighter bound



- Same objective: drawing cones
- New, tighter, asymmetric bounds
- Given
  - all definitions
  - NB: ε_1 the sub-optimality

Then for all $t' < t$:

$$F(e') \leq F(e) + \Phi_e\varepsilon_1 + \Phi_e \cdot (e_2 - e_1 - M_{max})(t' - t),$$

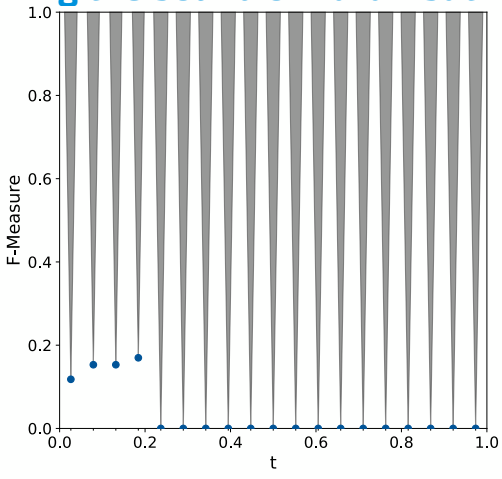
where $M_{max} = \max_{\substack{e'' \in \mathcal{E}(\mathcal{H}) \\ s.t.\ F(e'') > F(e)}} (e''_2 - e''_1)$

and, for all $t' > t$:

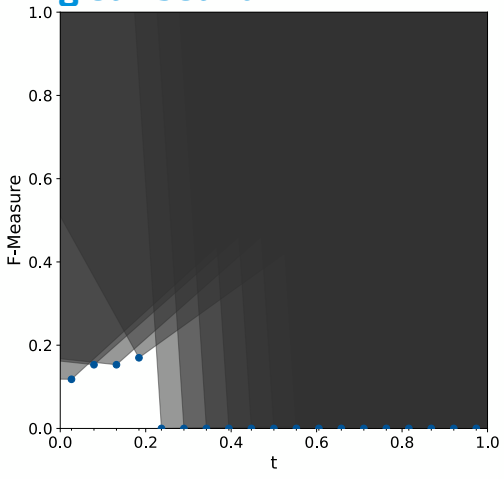
$$F(e') \leq F(e) + \Phi_e\varepsilon_1 + \Phi_e \cdot (e_2 - e_1 - M_{min})(t' - t),$$

where $M_{min} = \min_{\substack{e'' \in \mathcal{E}(\mathcal{H}) \\ s.t.\ F(e'') > F(e)}} (e''_2 - e''_1).$

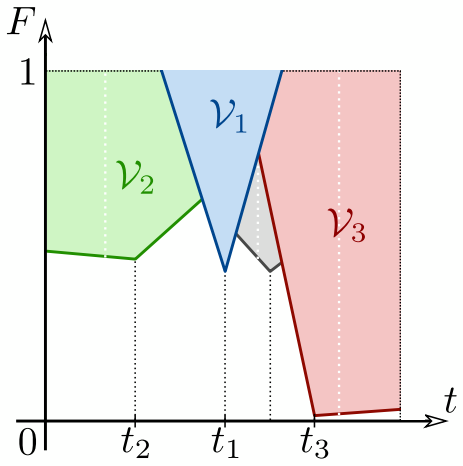
Visualizing the bound of Parambath et al.



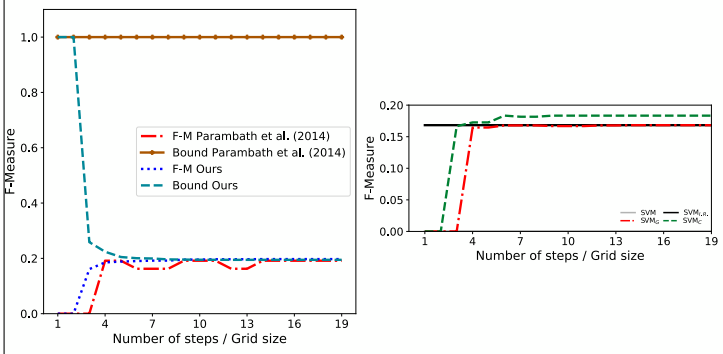
Visualizing our bound



Using the bound to better explore t values



CONE: example results



Non-vacuous bounds, incremental algo., early release.

CONE: numerical results

Dataset	SVM	$SVM_{I.R.}$	SVM_G	SVM_C	SVM_O^T	LR^T	$LR_{I.R.}^T$	LR_O^T	LR_B
Adult	62.5 (0.2)	64.9 (0.3)	66.4 (0.1)	66.5 (0.1)	66.4 (0.1)	66.5 (0.1)	66.5 (0.1)	66.5 (0.1)	**66.6** (0.1)
Abalone10	0.0 (0.0)	30.9 (1.2)	**32.4** (1.3)	32.2 (0.8)	31.8 (1.9)	30.8 (2.2)	30.7 (1.9)	30.7 (1.9)	31.6 (0.6)
Satimage	0.0 (0.0)	23.4 (4.3)	20.4 (5.3)	20.6 (5.6)	**30.9** (2.0)	21.2 (11.1)	28.6 (1.9)	28.6 (1.9)	21.4 (4.6)
IJCNN	44.5 (0.4)	53.3 (0.4)	61.6 (0.6)	61.6 (0.6)	**62.6** (0.4)	59.4 (0.5)	56.5 (0.3)	56.5 (0.3)	59.2 (0.3)
Abalone12	0.0 (0.0)	16.8 (2.7)	16.8 (4.2)	**18.3** (3.3)	16.3 (3.0)	15.5 (3.1)	17.0 (3.3)	17.0 (3.3)	17.7 (3.7)
Pageblocks	48.1 (5.8)	39.6 (4.7)	66.4 (3.2)	62.8 (3.9)	**67.6** (4.0)	59.2 (8.1)	55.9 (6.4)	55.9 (6.4)	55.7 (5.7)
Yeast	0.0 (0.0)	29.4 (2.9)	38.6 (7.1)	39.0 (7.5)	35.4 (15.6)	37.4 (10.1)	**39.9** (6.5)	27.6 (6.8)	27.6 (6.8)
Wine	0.0 (0.0)	15.6 (5.2)	20.0 (6.4)	22.7 (6.0)	19.3 (7.9)	21.5 (3.7)	**25.2** (4.5)	**25.2** (4.5)	18.3 (7.2)
Average	19.4 (0.8)	34.2 (2.8)	40.3 (3.5)	40.5 (3.5)	**41.3** (4.4)	38.9 (5.2)	40.0 (3.1)	38.5 (3.2)	37.3 (3.6)

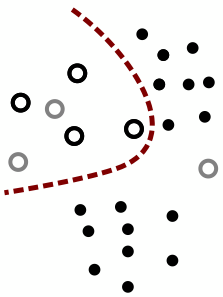
F-Measure for Logistic Regression (LR) [3] and SVM algorithms (averaged over 5 experiments).
"$I.R.$" I.R. based class costs; "G": Reproduction of [1]; "C": CONE "T" thresholded predictions; "B": Bissection [2]

No huge gain, better -measure in average.

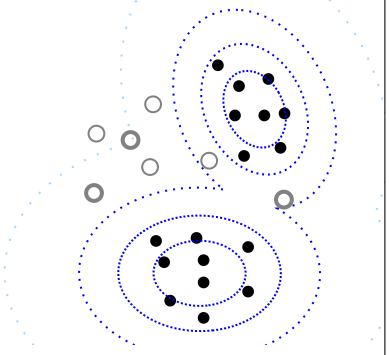
Overview

Reminder: Class Imbalance or Novelty?

Supervised
(some anomalies)



Unsupervised
(only normal points)



Overview

Simple (but realistic) unsupervised anomaly detection

- Setup and approach: *the three-sigma rule*
  - we have a set of (unlabeled) points
  - we consider one feature of interest
  - we look at the **standard deviation** σ and mean of this feature
  - anything beyond 3σ is an outlier/anomaly
- Possible improvements
  - use robust statistics (percentiles, robust estimation)
  - use several features

Process behind the 3σ rule

- Overall process
  - we suppose a **parametric model that explains the data**
    - i.e., we suppose the data is generated by this model
    - here, the data comes from the a normal distribution (with two unknown parameters: a mean μ and a standard deviation σ)
  - we estimate the parameters from the data
    - here, using the empirical mean and stdev
  - if the likelihood of a (new) point is low, it is an outlier
    - here, the normal density is very low after 3σ

Overview

- Introduction
- Anomaly and (rare) event detection
  - Problem, notations and performance measures
- Imbalanced classification problems
  - General approaches
  - Correcting k-NN: γ-***NN*** and ***MLFP***
  - Learning Maximum Excluding Ellipsoids
  - Focusing on the F-Measure optimization
- Unsupervised anomaly detection
  - Simple motivating approach
  - A variety of ML methods
  - Probabilistic models 101
  - Case study: temporal motif mining
  - More models (VAE, GAN, tensor networks, ...)
- Closing remarks

A few methods (ex. from scikit-learn...)

Overview

- Introduction
- Anomaly and (rare) event detection
  - Problem, notations and performance measures
- Imbalanced classification problems
  - General approaches
  - Correcting k-NN: γ-***NN*** and ***MLFP***
  - Learning Maximum Excluding Ellipsoids
  - Focusing on the F-Measure optimization
- Unsupervised anomaly detection
  - Simple motivating approach
  - A variety of ML methods
  - Probabilistic models 101
  - Case study: temporal motif mining
  - More models (VAE, GAN, tensor networks, ...)
- Closing remarks

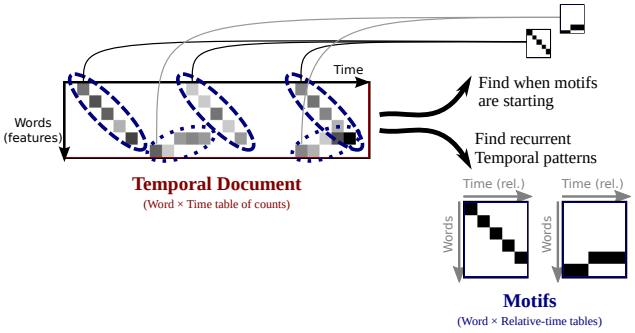
Probabilistic Generative Models

- A generalized "three sigma" rule
  - more complex models
  - more application than just anomaly detection
- Modeling step
  - we define a (stochastic) generative story
  - we define how we suppose data are generated
  - we encode our knowledge/assumptions/constraints
  - we define what is random and what is a parameter
- Learning/fitting step
  - given the data, what do we know about parameters
  - given the data, we find *the best* parameters
- Some possible usage
  - given the learned parameters, what is an outlier?
  - what do the parameters look like?
  - what data could I generate from these parameters?

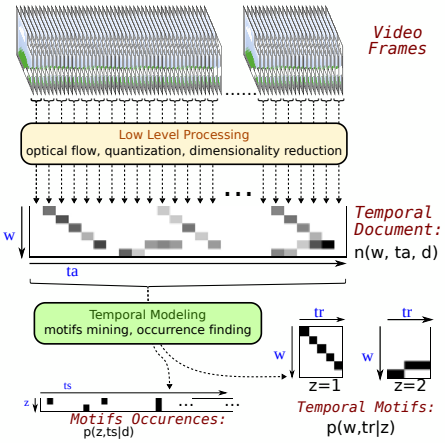
Overview

- Introduction
- Anomaly and (rare) event detection
  - Problem, notations and performance measures
- Imbalanced classification problems
  - General approaches
  - Correcting k-NN: γ-***NN*** and ***MLFP***
  - Learning Maximum Excluding Ellipsoids
  - Focusing on the F-Measure optimization
- Unsupervised anomaly detection
  - Simple motivating approach
  - A variety of ML methods
  - Probabilistic models 101
  - Case study: temporal motif mining
  - More models (VAE, GAN, tensor networks, ...)
- Closing remarks

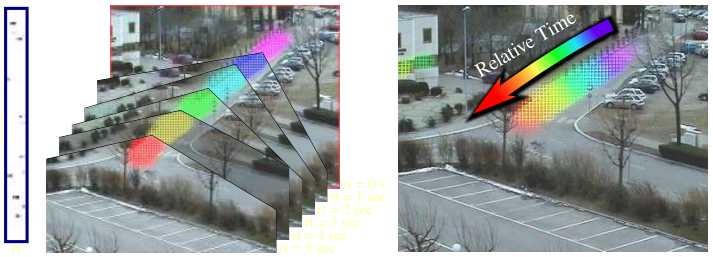
Unsupervised Temporal Motif Mining
in videos / temporal data (spectrograms, ...)



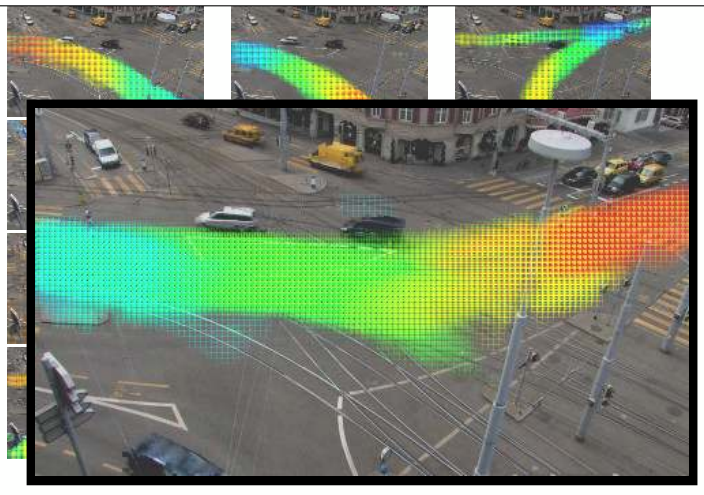
Temporal Patterns in Videos: Full Process



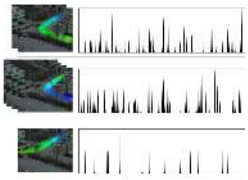
Video Motif Representation



Example Motifs Obtained from a Static Camera



Application with Static Cameras



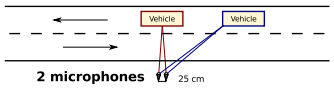
scene understanding

car counting

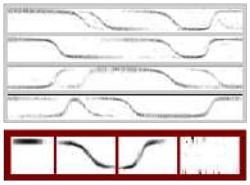
anomaly detection

stream selection, anomaly detection, multi-camera analysis

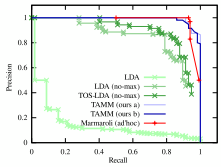
Audio data?



a pair of microphones…



… meaningful motifs…



… and good counting results

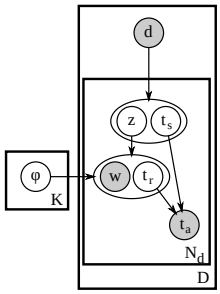
… also with spectrograms

How Can We Do This?!

Sol. 1: Hierarchical Probabilistic Models

$$\mathcal{L} = \sum_d \sum_w \sum_{t_a} n(w, t_a, d) log \sum_z \sum_{t_s} p(w, t_r|z) p(z, t_s|d)$$

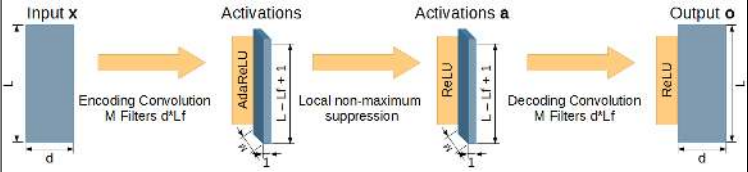
- Generative Model
  ⇒ interpretable by design
- ⚠ unknown number of motifs
  ⇒ use infinite models
- Inference
  - maximum likelihood, EM like
- Sparsity on occurrences $p(ts|z,d)$
  - new objective function:
    $\mathcal{L} - \lambda_{sparse} \sum_d \sum_z KL(U||p(ts|z,d))$



Sol. 2: Neural Networks, Auto-encoders

- Principle of auto-encoders
  - learn to produce the input from the input
  - going through a limited-size representation (bottleneck)
  - $x' = f(x) = f_{DEC}(f_{ENC}(x))$
  - minimize the reconstruction error $d(x, x') = \|x - x'\|^2$
- ⚠ Issues: interpretability, number of motifs, …

Sol. 2: Interpretable Auto-Encoders



- Add specific operators (layers)
  - global specialized maximum selection (AdaReLU)
  - locally, filter response de-correlation
- Special Loss: a combination of well-chosen target functions
  - encourage sparse motifs (with a lot of zero)
  - encourage sparse activations
  - ⚠ unknown number of motifs ⇒ use "group-sparsity"

Overview

- Introduction
- Anomaly and (rare) event detection
  - Problem, notations and performance measures
- Imbalanced classification problems
  - General approaches
  - Correcting k-NN: γ-***NN*** and ***MLFP***
  - Learning Maximum Excluding Ellipsoids
  - Focusing on the F-Measure optimization
- Unsupervised anomaly detection
  - Simple motivating approach
  - A variety of ML methods
  - Probabilistic models 101
  - Case study: temporal motif mining
  - More models (VAE, GAN, tensor networks, ...)
- Closing remarks

Overview

- Introduction
- Anomaly and (rare) event detection
  - Problem, notations and performance measures
- Imbalanced classification problems
  - General approaches
  - Correcting k-NN: γ-***NN*** and ***MLFP***
  - Learning Maximum Excluding Ellipsoids
  - Focusing on the F-Measure optimization
- Unsupervised anomaly detection
  - Simple motivating approach
  - A variety of ML methods
  - Probabilistic models 101
  - Case study: temporal motif mining
  - More models (VAE, GAN, tensor networks, ...)
- Closing remarks

Anomaly Detection with Tensor Networks

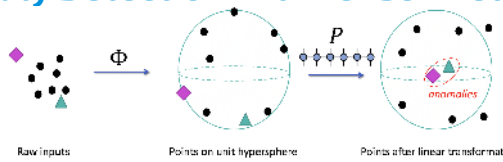


Figure 1: Schematic of Tensor Network Anomaly Detector (TNAD)

$$\|P\Phi(x)\|_2^2 =$$

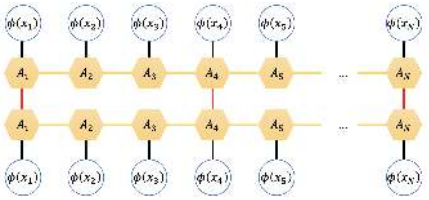
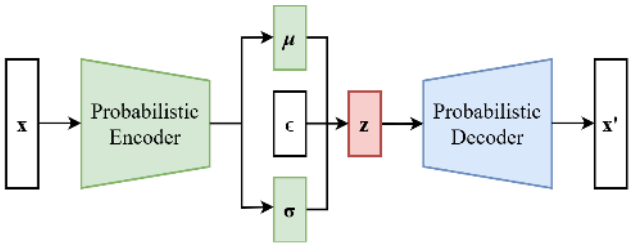
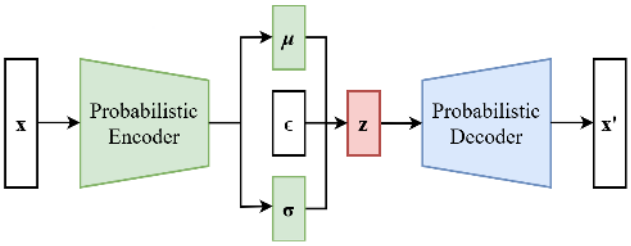


Figure 4: Squared norm of transformed vector.

Variational Autoencoders



GAN (for anomalies)

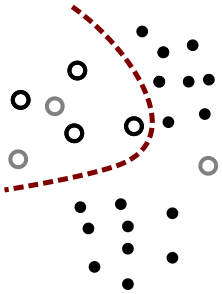


Overview

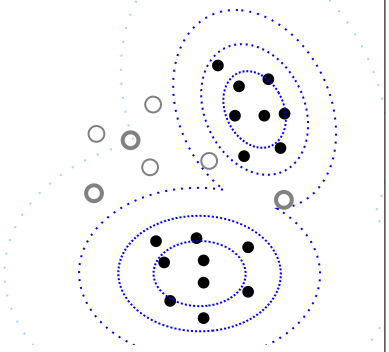
- Introduction
- Anomaly and (rare) event detection
  - Problem, notations and performance measures
- Imbalanced classification problems
  - General approaches
  - Correcting k-NN: γ-***NN*** and ***MLFP***
  - Learning Maximum Excluding Ellipsoids
  - Focusing on the F-Measure optimization
- Unsupervised anomaly detection
  - Simple motivating approach
  - A variety of ML methods
  - Probabilistic models 101
  - Case study: temporal motif mining
  - More models (VAE, GAN, tensor networks, ...)
- Closing remarks

Reminder: Class Imbalance or Novelty?

Supervised
(some anomalies)



Unsupervised
(only normal points)



Learning from Imbalanced Data and Anomaly Detection

Questions?

Rémi Emonet

Université Jean-Monnet,
Laboratoire Hubert Curien,
Saint-Étienne

Talk at SLEIGHT Science Event#6, 2021-07-06