Renyi divergence learning for explainable classification

Matthieu Gallet, **Ammar Mian**, Abdourrahmane Atto

ICASSP 2024

18 April 2024

LISTIC UNIVERSITÉ SAVOIE MONT BLANC METEO FRANCE ICASSP 2024 KOREA

Table of contents

Outline

Context:

Synthetic Aperture RADAR (SAR)

- Active sensor
- All-weather, day and night
- Complex labelling
- High clutter and geometric distortions
- Multiplicative noise

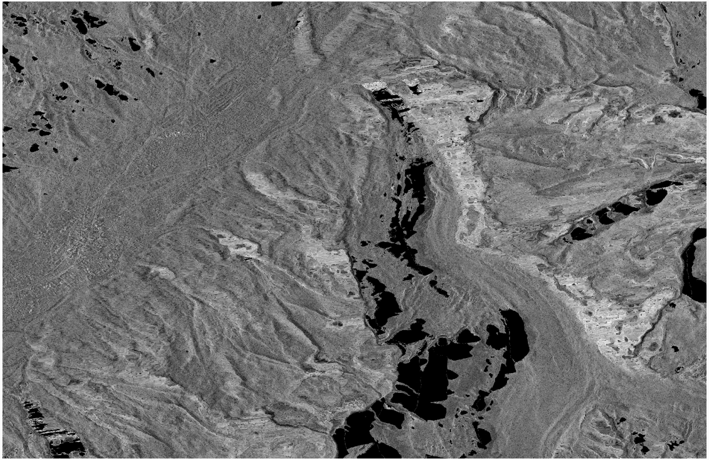
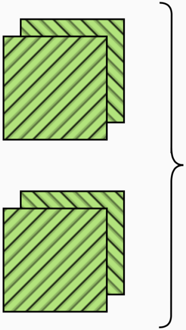


Figure 1: Acquisition in X-band the 10th January 2020.

Motivation

Image 1

Image 2

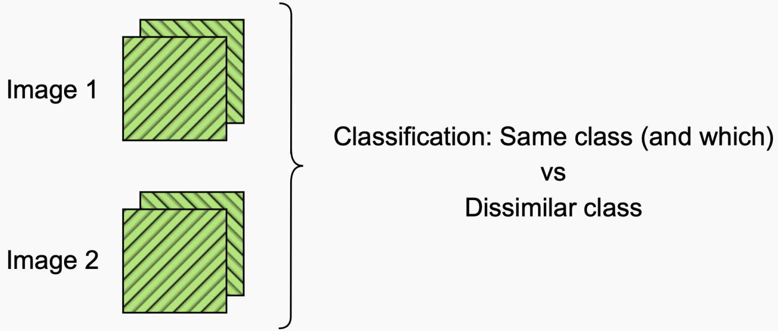


Classification: Same class (and which)

vs

Dissimilar class

Motivation



- Different sensors, acquisition modalities
- Can be general but we consider case bivariate: two polarizations

Context:

Motivation: Pairwise classification of SAR images, bivariate.

Context:

Motivation: Pairwise classification of SAR images, bivariate.

- **Statistically based**: Based on divergences [Cilingir et al., 2020], Wishart distributions [Silva et al., 2013]
  - Pro: Explainability, small amount of data
  - Cons: Model assumption, quality of estimation

Context:

Motivation: Pairwise classification of SAR images, bivariate.

- **Statistically based**: Based on divergences [Cilingir et al., 2020], Wishart distributions [Silva et al., 2013]
  - Pro: Explainability, small amount of data
  - Cons: Model assumption, quality of estimation
- **Model free**: deep-learning, features extraction [Ansari et al., 2020] [Chen et al., 2016]
  - Pro: No model assumption, data driven
  - Cons: Lack of explainability, need a large amount of data or good labels, dealing with the multiplicative noise of SAR images

Motivation: Pairwise classification of SAR images, bivariate.

- **Statistically based**: Based on divergences [Cilingir et al., 2020], Wishart distributions [Silva et al., 2013]
  - Pro: Explainability, small amount of data
  - Cons: Model assumption, quality of estimation
- **Model free**: deep-learning, features extraction [Ansari et al., 2020] [Chen et al., 2016]
  - Pro: No model assumption, data driven
  - Cons: Lack of explainability, need a large amount of data or good labels, dealing with the multiplicative noise of SAR images

Proposition

Combination of both approaches: using multiple proability models to extract features (parameters of model) and to combine them using a combination-metric learned from the data.

Outline

Parametric SAR model and features 1/2

Let $\mathbf{I}_{i,j}$ be a pair of a j-variate patch of SAR images, iid. We construct the pair of vector of features $\hat{\mathbf{x}}_{i,j}$ as follows:

$$\hat{\mathbf{x}}_{i,j} = \left[\theta_{\mathcal{G}}(\mathbf{I}_{i,j}), \theta_{\mathcal{O}}(\mathbf{I}_{i,j}), \theta_{\mathcal{R}}(\mathbf{I}_{i,j})\right]^{\mathrm{T}}, \forall i, j$$

where $\theta_{\mathcal{G}}(\mathbf{I}_{i,j})$, $\theta_{\mathcal{O}}(\mathbf{I}_{i,j})$ and $\theta_{\mathcal{R}}(\mathbf{I}_{i,j})$ are the parameters of the three following distributions fitted on the amplitude of the SAR patch $\mathbf{I}_{i,j}$.

Features

Gamma: $\mathcal{G}(x;\mu,L) = e^{-\frac{xL}{\mu}} \cdot \left(\frac{L}{\mu}\right)^{L} \cdot \Gamma(L) \cdot x^{L-1}$, with shape and scale L and μ,

log-normal: $\mathcal{O}(x;\mu,\sigma) = e^{-\frac{(\log x-\mu)^2}{2\sigma^2}} \cdot \frac{1}{x\sigma\sqrt{2\pi}}$, with mean μ and variance σ,

Rayleigh: $\mathcal{R}(x;\mu) = \frac{x}{2\mu^2} \cdot e^{-\left(\frac{x}{2\mu}\right)^2}$ with scale μ,

Parametric SAR model and features 2/2

From:

$$\hat{\mathbf{x}}_{i,j} = \left[\theta_{\mathcal{G}}(\mathbf{I}_{i,j}), \theta_{\mathcal{O}}(\mathbf{I}_{i,j}), \theta_{\mathcal{R}}(\mathbf{I}_{i,j})\right]^{\mathrm{T}}, \forall i, j$$

We have:

$$\mathbf{X} = (\mathbf{x}_1, \mathbf{x}_2) = \begin{pmatrix} \hat{\mathbf{x}}_{1,1} & \hat{\mathbf{x}}_{2,1} \\ \vdots & \vdots \\ \hat{\mathbf{x}}_{1,J} & \hat{\mathbf{x}}_{2,J} \end{pmatrix}$$

We consider in the following the bivariate case ($J = 2$).

Divergences i

Divergences

The Rényi divergence of order α between two probability distributions P, Q on $\mathbb{R}^n$ is given by:

$$D_\alpha(P\|Q) = \frac{1}{\alpha - 1} \ln \int P(x)^\alpha Q(x)^{1-\alpha} \, dx, \tag{1}$$

with $\alpha > 0$ and $\alpha \neq 1$.

Closed form for the Gamma, log-normal and Rayleigh distributions [Gil et al., 2013]:

Divergences ii

- Gamma:

$$D_\alpha(P\|Q) = \ln\left(\frac{\Gamma\left(k_j\right)\theta_j^{k_j}}{\Gamma\left(k_i\right)\theta_i^{k_i}}\right) + \frac{1}{\alpha-1}\ln\left(\frac{\Gamma\left(k_\alpha\right)}{\theta_i^{k_i}\Gamma\left(k_i\right)}\left(\frac{\theta_i\theta_j}{\theta_\alpha^*}\right)^{k_\alpha}\right)$$

$$\theta_\alpha^* = \alpha\theta_j + (1-a)\theta_i, k_\alpha = \alpha k_i + (1-\alpha)k_j$$

$$\theta_\alpha^* > 0 \text{ and } k_\alpha > 0$$

Divergences iii

- log-normal:

$$D_\alpha(P\|Q) = \ln\frac{\sigma_j}{\sigma_i} + \frac{1}{2(\alpha-1)}\ln\left(\frac{\sigma_j^2}{(\sigma^2)^*_\alpha}\right) + \frac{1}{2}\frac{\alpha\left(\mu_i-\mu_j\right)^2}{(\sigma^2)^*_\alpha}$$

$$\left(\sigma^2\right)^*_\alpha = \alpha\sigma_j^2 + (1-\alpha)\sigma_i^2$$

$$\left(\sigma^2\right)^*_\alpha > 0$$

- Rayleigh:

$$D_\alpha(P\|Q) = 2\ln\frac{\sigma_j}{\sigma_i} + \frac{1}{\alpha-1}\ln\left(\frac{\sigma_j^2}{(\sigma^2)^*_\alpha}\right)$$

$$\left(\sigma^2\right)^*_\alpha = \alpha\sigma_j^2 + (1-\alpha)\sigma_i^2$$

$$\left(\sigma^2\right)^*_\alpha > 0$$

Pipeline: non parametric 1/2

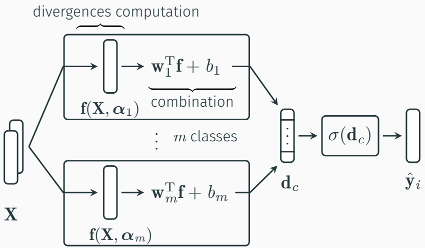


Figure 2: Schema of the pipeline.

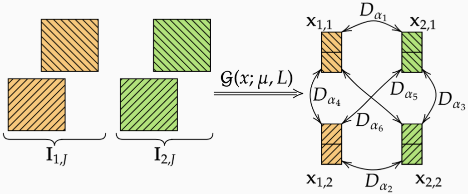


Figure 3: Diagram of the divergence estimation for one distribution.

Let $\mathbf{f} = \left[D_{\alpha_1}(\mathbf{x}_1^1, \mathbf{x}_2^1), ..., D_{\alpha_p}(\mathbf{x}_1^p, \mathbf{x}_2^p)\right]^{\mathrm{T}}$ be a vector composed by a set of p Renyi divergences with parameters $\boldsymbol{\alpha} = [\alpha_1, \ldots, \alpha_p]^{\mathrm{T}} \in (0,1)^p$.
For 3 distributions considered, the number of divergences is $p = 3 \times \binom{i \times j}{2}$

Pipeline: non parametric 2/2

For each class c, given a set of parameters $\boldsymbol{\alpha}_c$ we combine the divergences:

$$\mathbf{d}_c(\mathbf{X}, \boldsymbol{\alpha}_c) = \mathbf{w}_c^{\mathrm{T}} \mathbf{f}(\mathbf{X}, \boldsymbol{\alpha}_c) + b_c, \tag{2}$$

where $\mathbf{w}_c \in \mathbb{R}_+^p$ and $b_c \in \mathbb{R}_+$.

Explainability

- p parameters $\boldsymbol{\alpha}_c$ for each class c.
- constraints on $\boldsymbol{\alpha}_c \in (0, 1)$.
  - $\boldsymbol{\alpha}_c \to 1$, $D_\alpha(P\|Q) \to \mathrm{KL}(P\|Q)$.
  - $\boldsymbol{\alpha}_c \to 0.5$, $D_\alpha(P\|Q)$ homogeneous to Bhattacharyya distance.
- positive constraints on $\mathbf{w}_c$ and b_c.

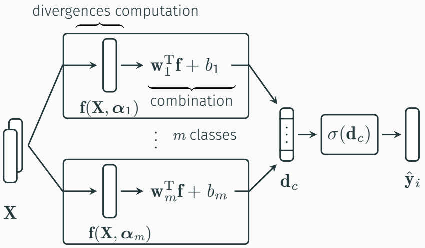


Figure 4: Schema of the pipeline.

Close to [Cilingir et al., 2020], but more constrained.

Minimization problem

We consider the cross-entropy loss: $H(y,\hat{y}) = -\sum_c^m y_c \log(\hat{y}_c)$, with $\hat{y}, y \in \mathbb{R}^m$ (the prediction of a classifier and its associated ground truth) and σ the softmax function. This gives us the following minimization problem:

$$\operatorname{argmin}_{\substack{\forall c \in \{1,\ldots,m\},\\ \boldsymbol{\alpha}_c \in (0,1)^p,\\ \mathbf{w}_c \in \mathbb{R}_+^p, b_c \in \mathbb{R}_+}} \frac{1}{n} \sum_{i=0}^{n} \underbrace{-\sum_{c}^{m} \mathbf{y}_i(c) \log\left[\sigma \circ \mathbf{d}_c(\mathbf{X}_i, \boldsymbol{\alpha}_c)\right]}_{\mathcal{L}_i}. \tag{3}$$

Optimization

b_c and $\mathbf{w}_c$ are updated with a standard gradient descent.

we provide a closed form for the gradient of $\boldsymbol{\alpha}_c$.

Gradient of α_c

Divergence	**derivate** $\partial D_{\hat{\alpha}_{(\cdot)}} / \partial \alpha_{(\cdot)}$
$D_{\hat{\alpha}_l}(\mathcal{G}_i \| \mathcal{G}_j)$	$e^{\alpha_l} \frac{L_i \mu_j - L_j \mu_i}{\lambda_{ij} \beta_{ij}} - e^{\alpha_l} L_i \log \left[\frac{(e^{\alpha_l}+1)\mu_i \mu_j}{\beta_{ij}} \right] - e^{\alpha_l} \log \left[\frac{\left(\frac{\mu_i}{L_i}\right)^{-L_i} \Gamma(\lambda_{ij})}{\Gamma(L_i)} \right] + e^{\alpha_l} \frac{(L_j - L_i)\psi^{(0)}(\lambda_{ij})}{e^{\alpha_l}+1}$
$D_{\hat{\alpha}_l}(\mathcal{O}_i \| \mathcal{O}_j)$	$\frac{e^{\alpha_l}}{2(e^{\alpha_l}\Sigma_{ij})^2} \left[e^{\alpha_l} \sigma_j^2 (\sigma_j^2 - \sigma_i^2) + \sigma_i^2 \left[(\mu_i - \mu_j)^2 - \Sigma_{ij} \right] - (e^{\alpha_l} \Sigma_{ij})^2 \log \left(\frac{(e^{\alpha_l}+1)\sigma_j^2}{e^{\alpha_l}\Sigma_{ij}} \right) \right]$
$D_{\hat{\alpha}_l}(\mathcal{R}_i \| \mathcal{R}_j)$	$\frac{\gamma_{ij} - e^{\alpha_l}\mu_i^2}{\gamma_{ij} + \mu_i^2} - e^{\alpha_l} \log \left[\frac{\gamma_{ij} + \mu_j^2}{\gamma_{ij} + \mu_i^2} \right]$

Table 1: Rényi's derivate

with $\hat{\alpha} = 1/(1 - \mathrm{e}^{-\alpha})$ and:

$$\lambda_{ij} = (e^{\alpha_l} L_i + L_j)/(1 + e^{\alpha_l}),$$
$$\beta_{ij} = e^{\alpha_l} L_i \mu_j + L_j \mu_i,$$
$$\Sigma_{ij} = \sigma_j^2 + \sigma_i^2,$$
$$\gamma_{ij} = e^{\alpha_l} \mu_j^2.$$

Outline

Results

X-band SAR dataset dual-pol (HH, HV)

645 patches of 32x32 pixels

5 classes (glacier, city, forest, rock, plain) + 1 class for the dissimilar

comparison with a CNN [Parikh et al., 2020] and a Random Forest (RF)

for a bivariate pair with 3 distributions: $3 \times \binom{4}{2}$=18 divergences per class

	RF	CNN	Rényi
input size	200×1	$32 \times 32 \times 4$	10×2
parameters	$\sim 152,000$	$226,406$	222

Table 2: Number of parameters and size of the inputs used

	ACC	CIT	DIS	FOR	PLA	ROC
RF	65.3 ± 13.0	70.8 ± 9.2	82.7 ± 4.1	11.9 ± 1.2	33.6 ± 5.8	55.9 ± 9.0
CNN	$\mathbf{83.5 \pm 7.0}$	61.1 ± 16.5	$\mathbf{82.9 \pm 4.3}$	45.1 ± 8.1	49.5 ± 13.0	$\mathbf{72.3 \pm 1.2}$
Renyi	59.1 ± 11.1	$\mathbf{83.2 \pm 4.2}$	45.3 ± 1.2	$\mathbf{80.5 \pm 6.9}$	$\mathbf{67.3 \pm 3.7}$	62.7 ± 12.0

Table 3: Percentage of good classification with a stratified K-Fold with K=5

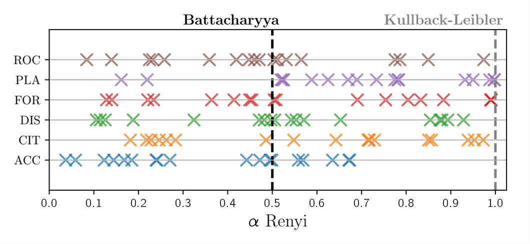


Figure 5: α learned for each features and for each classes.

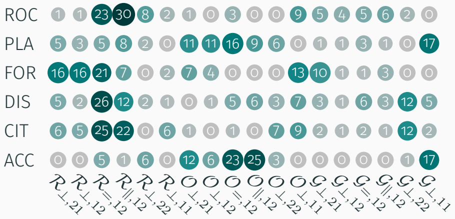


Figure 6: Visualisation of associated weights in the decision process for each class.

Robustness to noise

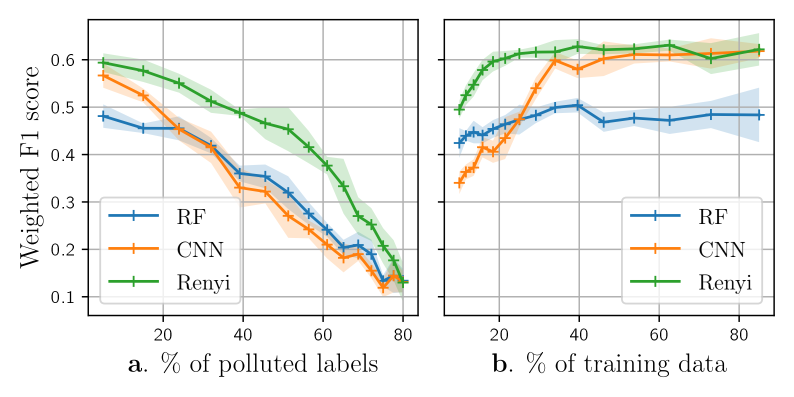


Figure 7: Comparison of performance (mean of weighted f1 score over all class in function) of two1 perturbations, **a.** Percentage of label perturbation and **b.** Percentage of data training

Outline

Conclusion

What we have done

- New solution by joint use of parametric and non-parametric methods
- Derivate the analytical gradient for three distributions wrt the Renyi parameter (learning)
- Less parameters than traditional ML methods
- Explainability of the classification and robustness to noise

What's next?

- Treat the case $\alpha > 1$ and find a solution for $\alpha = 1$.
- Consider different distributions between pairs
- Study convergence of gradient descent
- Metric learning problems

References i

Ansari, R. A., Buddhiraju, K. M., and Malhotra, R. (2020).
Urban change detection analysis utilizing multiresolution texture features from polarimetric sar images.
Remote Sensing Applications: Society and Environment, 20:100418.

Chen, S., Wang, H., Xu, F., and Jin, Y.-Q. (2016).
Target classification using the deep convolutional networks for sar images.
IEEE Transactions on Geoscience and Remote Sensing, 54(8):4806–4817.

Cilingir, H. K., Manzelli, R., and Kulis, B. (2020).
Deep divergence learning.
In *International Conference on Machine Learning*, pages 2027–2037. PMLR.

References ii

Gil, M., Alajaji, F., and Linder, T. (2013).
Rényi divergence measures for commonly used univariate continuous distributions.
Information Sciences, 249:124–131.

Parikh, H., Patel, S., and Patel, V. (2020).
Classification of SAR and PolSAR images using deep learning: a review.
International Journal of Image and Data Fusion, 11(1):1–32.
Number: 1.

Silva, W. B., Freitas, C. C., Sant'Anna, S. J., and Frery, A. C. (2013).
Classification of segments in polsar imagery by minimum stochastic distances between wishart distributions.
IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing, 6(3):1263–1273.