A Non-Convex Proximal Approach for Centroid-based Classification

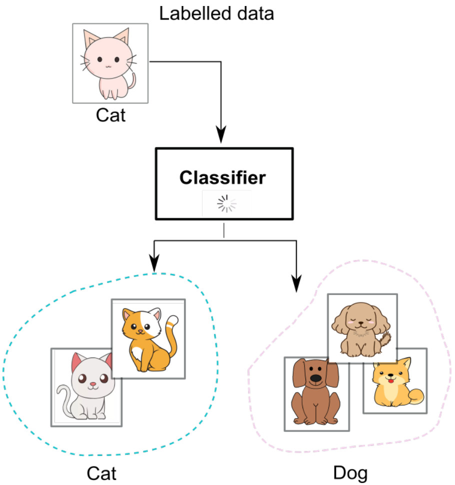
2022 IEEE International Conference on Acoustics, Speech and Signal Processing

Mewe-Hezoudah Kahanam[†] Laurent Le-Brusquet[⋆] Ségolène Martin[†]
Jean-Christophe Pesquet[†]

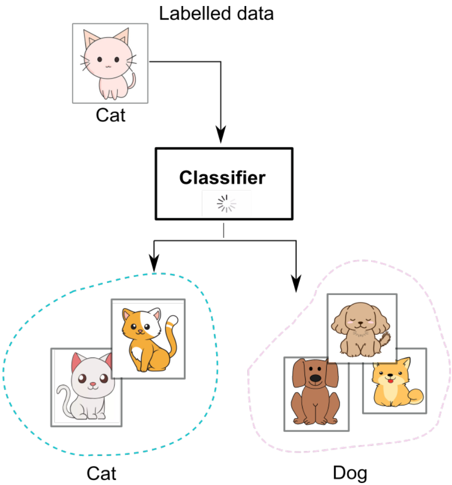
[†] Université Paris-Saclay, Inria, CentraleSupélec, Centre de Vision Numérique
[⋆] Université Paris-Saclay, CNRS, CentraleSupélec, Laboratoire des Signaux et Systèmes

May 10th 2022

Supervised classification



Supervised classification



Two widely used techniques in classification:

- Data dimension reduction
- Estimate the centroids of the classes

Main objective of this work

Propose a new method for **supervised classification method** relying on a minimization problem that couples:

- data **dimension reduction** through a linear transform,
- estimation of the **centroids** of the classes.

Problem's variables

Given inputs for the training:

- m samples, k classes,
- $\mathbf{X} \in \mathbb{R}^{m \times d}$ matrix containing the m samples,
- $\mathbf{Y} \in \{0,1\}^{m \times k}$ matrix of one-hot encoded labels.

Problem's variables

Given inputs for the training:

- m samples, k classes,
- $\mathbf{X} \in \mathbb{R}^{m \times d}$ matrix containing the m samples,
- $\mathbf{Y} \in \{0, 1\}^{m \times k}$ matrix of one-hot encoded labels.

Unknowns:

- linear transform $\mathbf{W} \in \mathbb{R}^{d \times \ell}$ with $\ell \ll d$,
- matrix of centroids $\mathbf{M} = [\mathbf{M}_1, \ldots \mathbf{M}_k]^\top \in \mathbb{R}^{k \times \ell}$.

Minimization problem

Centroid-based classification approach amounts to minimizing the following loss function with respect to the matrix variables **W** (transform) and **M** (centroids) :

$$\underset{(\mathbf{M},\mathbf{W})}{\text{minimize}} \quad f(\mathbf{YM} - \mathbf{XW}) + g(\mathbf{W}) + h(\mathbf{M})$$

where f is row-wise separable, i.e.

$$(\forall \mathbf{Z} = [\mathbf{Z}_1, \dots, \mathbf{Z}_m]^\top \in \mathbb{R}^{m\times\ell}) \quad f(\mathbf{Z}) = \sum_{i=1}^{m} \varphi(\mathbf{Z}_i),$$

for some function $\varphi \colon \mathbb{R}^\ell \longrightarrow]-\infty, +\infty]$.

☞ Once **W** and **M** were obtained using the training set, a new sample $\mathbf{X}_{m+1}$ can be assigned to a class j^*, where j^* satisfies

$$j^* \in \underset{j\in\{1,\dots,k\}}{\text{Argmin}} \; \varphi(\mathbf{M}_j - \mathbf{W}^\top \mathbf{X}_{m+1}).$$

Minimization problem

Centroid-based classification approach amounts to minimizing the following loss function with respect to the matrix variables $\mathbf{W}$ (transform) and $\mathbf{M}$ (centroids):

$$\underset{(\mathbf{M},\mathbf{W})}{\text{minimize}} \quad f(\mathbf{YM} - \mathbf{XW}) + g(\mathbf{W}) + h(\mathbf{M})$$

Example:

- A classical choice for f corresponds to $f = \| \cdot \|_F^2$. In that case:

$$f(\mathbf{YM} - \mathbf{XW}) = \sum_{j=1}^{k} \sum_{i \in C_j} \|\mathbf{M}_j - \mathbf{W}^\top \mathbf{X}_i\|_2^2,$$

  where $(C_j)_{1 \leq j \leq k}$ denote the classes.
  Alternative choice: $f = \| \cdot \|_1$

Minimization problem

Centroid-based classification approach amounts to minimizing the following loss function with respect to the matrix variables **W** (transform) and **M** (centroids) :

$$\underset{(\mathbf{M},\mathbf{W})}{\text{minimize}} \quad f(\mathbf{YM} - \mathbf{XW}) + g(\mathbf{W}) + h(\mathbf{M})$$

Example:

- A sparsity-promoting regularization is often employed for **W**, corresponding to $g = \alpha\| \cdot \|_1$, where $\alpha > 0$.

Minimization problem

Centroid-based classification approach amounts to minimizing the following loss function with respect to the matrix variables $\mathbf{W}$ (transform) and $\mathbf{M}$ (centroids):

$$\underset{(\mathbf{M},\mathbf{W})}{\text{minimize}} \quad f(\mathbf{YM} - \mathbf{XW}) + g(\mathbf{W}) + h(\mathbf{M})$$

Example:

- Choice for function h ?

Centroid separation term

We opt for a particular choice of function h which encourages the separation of the centroids, namely

$$(\forall \mathbf{M} \in \mathbb{R}^{k\times\ell}) \quad h(\mathbf{M}) = -\gamma \sum_{1\leqslant i<j\leqslant k} \|\mathbf{M}_j - \mathbf{M}_i\|_1.$$

Note that h is **nonconvex**, which makes the optimization problem difficult to solve.

How to avoid trivial solutions?

⚠ Issue with this model:

One of the following is likely to happen for standard choices of functions f and g:

- the criterion is unbounded from below,
- $(\mathbf{M}, \mathbf{W}) = (\mathbf{0}, \mathbf{0})$ is a trivial solution.

How to avoid trivial solutions?

☞ we bound the centroid matrix $\mathbf{M}$ by constraining each of the centroids $(\mathbf{M}_j)_{1 \leq j \leq k}$ to lie in a closed ball of radius $\delta > 0$.
The modified minimization problem is:

$$\begin{aligned} & \underset{(\mathbf{M},\mathbf{W})}{\text{minimize}} \quad f(\mathbf{YM} - \mathbf{XW}) + g(\mathbf{W}) + h(\mathbf{M}) \\ & \text{subject to} \quad \mathbf{M} \in C \end{aligned} \tag{2}$$

where

$$C = \left\{ \mathbf{M} \in \mathbb{R}^{k \times \ell} \mid (\forall j \in \{1, \dots, k\}) \quad \|\mathbf{M}_j\|_2 \leq \delta \right\},$$

where $\delta > 0$ is a fixed parameter.

An equivalent formulation

Rewriting the ℓ_1-norm through its dual norm, we can define a matrix $\mathbf{A} \in \mathbb{R}^{(\ell(\ell-1)/2)\times k}$ such that

$$h(\mathbf{M}) = -\gamma \sum_{1\leqslant i<j\leqslant k} \|\mathbf{M}_j - \mathbf{M}_i\|_1 = -\gamma \max_{\|\mathbf{U}\|_\infty \leq 1} \langle \mathbf{AM}, \mathbf{U}\rangle,$$

Therefore, Problem (2) is equivalent to

$$\begin{aligned} &\underset{(\mathbf{M},\mathbf{W},\mathbf{U})}{\text{minimize}} \quad f(\mathbf{YM} - \mathbf{XW}) + g(\mathbf{W}) - \gamma\langle \mathbf{AM}, \mathbf{U}\rangle \\ &\text{subject to} \;\; \mathbf{M} \in C \text{ and } \|\mathbf{U}\|_\infty \leq 1 \end{aligned} \tag{3}$$

☞ The above problem is **convex with respect to each variable** $\mathbf{M}, \mathbf{W}$, and $\mathbf{U}$ when f and g are convex.

Alternating proximal algorithm

–> perform a proximal minimization step on each one of the variable M, W, U successively

Accelerated primal-dual algorithm

–> to compute the proximal operator when it is not closed-form

Alternating proximal algorithm

–> perform a proximal minimization step on each one of the variable M, W, U successively

Accelerated primal-dual algorithm

–> to compute the proximal operator when it is not closed-form

✓ when f and g are convex proper l.s.c, the algorithm is guaranteed to converge to a critical point of the objective.

Numerical experiments

☞ We evaluate the performance of our method on the KEEL dataset

		texture	sonar	pima	wdbc	banana	magic	satimage	titanic	bupa	AVG
Ours	Train	87.0	85.4	74.7	94.1	57.3	77.1	78.3	77.3	59.7	**76.8**
	Test	86.4	72.0	73.0	94.0	54.9	77.2	77.3	77.3	58.2	**74.5**
Barlaud et al.	Train	72.8	83.1	76.5	88.0	56.0	66.0	74.7	77.6	69.3	73.8
	Test	72.3	68.1	75.5	87.2	54.4	65.7	72.1	77.6	67.8	71.2
NCM	Train	74.5	72.7	73.4	93.9	57.7	77.1	78.7	75.4	60.0	73.7
	Test	73.7	70.2	72.8	93.7	57.4	76.9	78.4	74.6	60.0	73.1

Table: Classification rate of our method compared to the state-of-the-art.

References

[Alcalá-Fdez et al., 2011] Alcalá-Fdez, J., Fernández, A., Luengo, J., Derrac, J., García, S., Sánchez, L., and Herrera, F. (2011).
Keel data-mining software tool: data set repository, integration of algorithms and experimental analysis framework.
Journal of Multiple-Valued Logic & Soft Computing, 17.

[Barlaud et al., 2019] Barlaud, M., Chambolle, A., and Caillau, J.-B. (2019).
Robust supervised classification and feature selection using a primal-dual method.
arXiv preprint arXiv:1902.01600.

[Chambolle and Pock, 2011] Chambolle, A. and Pock, T. (2011).
A first-order primal-dual algorithm for convex problems with applications to imaging.
Journal of Mathematical Imaging and Vision, 40(1):120–145.

[Mensink et al., 2013] Mensink, T., Verbeek, J., Perronnin, F., and Csurka, G. (2013).
Distance-based image classification: generalizing to new classes at near-zero cost.
IEEE Transactions on Pattern Analysis and Machine Intelligence, 35(11):2624–2637.