Adapting supervised classification algorithms to weak label scenarios

Miquel Perelló-Nieto[1], Raúl Santos-Rodríguez[1], and Jesús Cid-Sueiro[2]

[1]University of Bristol, UK
[2]Universidad Carlos III de Madrid, Spain

Email: [1]{Miquel.PerelloNieto, enrsr}@bristol.ac.uk, [2]jcid@tsc.uc3m.es

October 26, 2017

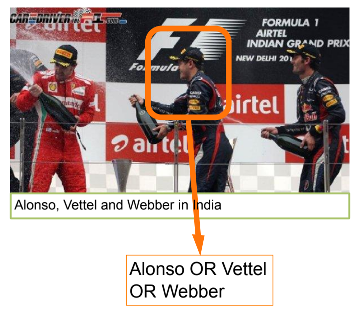
Index

Index

Index

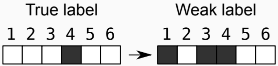
Weak labels: E.g. Captions

- One instance is one person
- Caption contains multiple labels (names)
- For every instance we have *multiple labels*
- Only one label is true

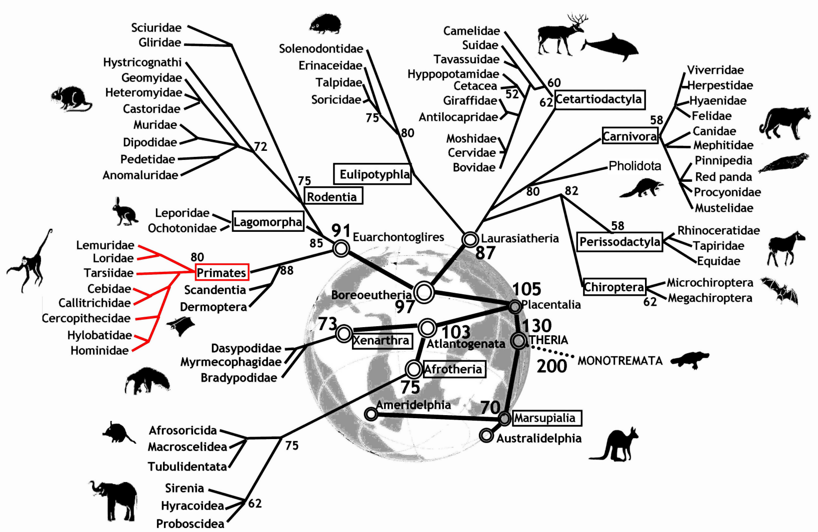


Alonso, Vettel and Webber in India

Alonso OR Vettel OR Webber

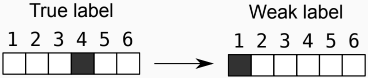


Weak labels: E.g. Supergroups



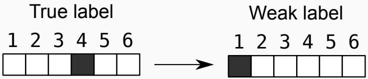
Weak labels: Other reasons

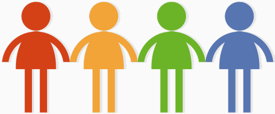
- Ambiguity
- Lack of expertise
- Cheaper annotations
- Crowdsourcing
  - Fast to obtain labels
  - People may disagree

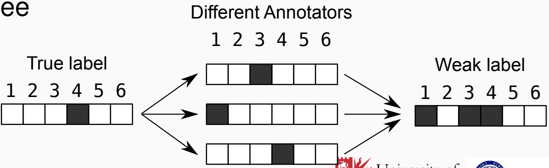


Weak labels: Other reasons

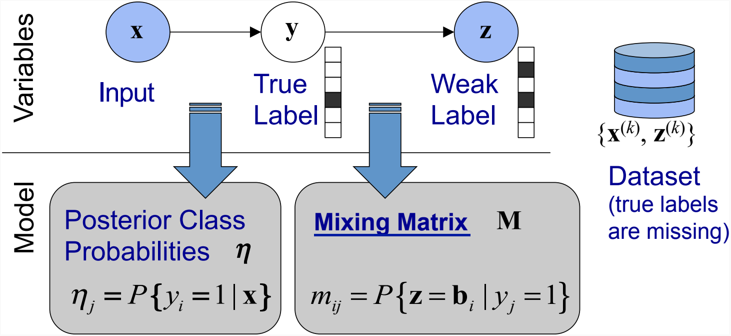
- Ambiguity
- Lack of expertise
- Cheaper annotations
- Crowdsourcing
  - Fast to obtain labels
  - People may disagree





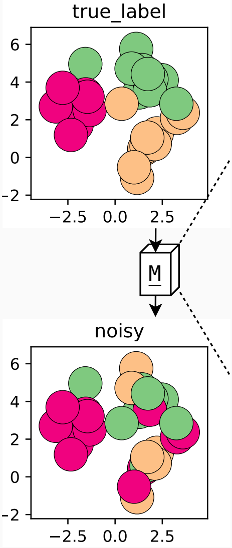


Data model


Variables
x
y
z
Input
True Label
Weak Label
$\{\mathbf{x}^{(k)}, \mathbf{z}^{(k)}\}$
Dataset (true labels are missing)
Model
Posterior Class Probabilities η
$\eta_j = P\{y_i = 1 \mid \mathbf{x}\}$
Mixing Matrix $\mathbf{M}$
$m_{ij} = P\{\mathbf{z} = \mathbf{b}_i \mid y_j = 1\}$

Index

Mixing matrices: Noisy labels



True Label

Weak Label

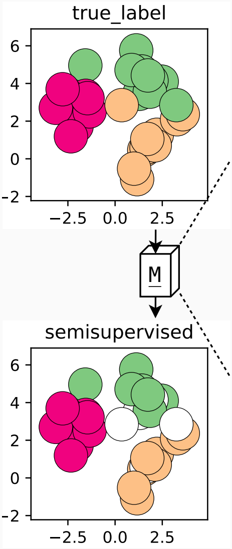
	■□□	□■□	□□■
□□□	0	0	0
■□□	$1-\beta$	$\beta/2$	$\beta/2$
□■□	$\beta/2$	$1-\beta$	$\beta/2$
□□■	$\beta/2$	$\beta/2$	$1-\beta$
■■□	0	0	0
■□■	0	0	0
□■■	0	0	0
■■■	0	0	0

$\sum x_i = 1$

Probability of observing weak label ■□■ (classes 1 and 3) when the true class is □□■ (class 3)

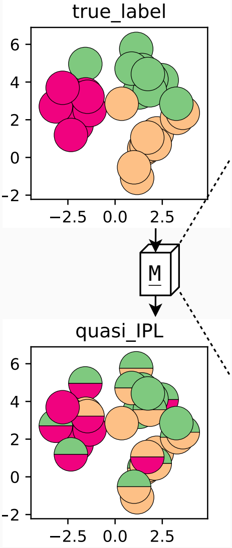
University of BRISTOL Universidad Carlos III de Madrid

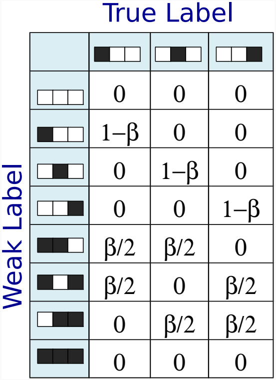
Mixing matrices: Semi-supervised learning



	True Label		
Weak Label	■□□	□■□	□□■
□□□	α	β	γ
■□□	1−α	0	0
□■□	0	1−β	0
□□■	0	0	1−γ
■■□	0	0	0
■□■	0	0	0
□■■	0	0	0
■■■	0	0	0

Mixing matrices: True label + noisy label

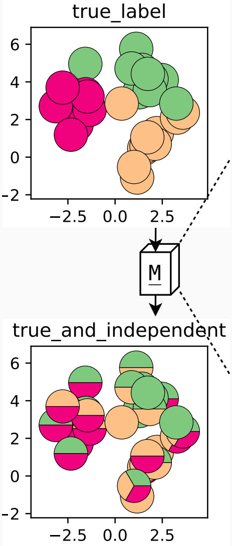



True Label

Weak Label

	■□□	□■□	□□■
□□□	0	0	0
■□□	$1-\beta$	0	0
□■□	0	$1-\beta$	0
□□■	0	0	$1-\beta$
■■□	$\beta/2$	$\beta/2$	0
■□■	$\beta/2$	0	$\beta/2$
□■■	0	$\beta/2$	$\beta/2$
■■■	0	0	0

Mixing matrices: True label + independent noisy label



True Label

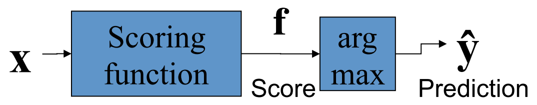
Weak Label	■□□	□■□	□□■
□□□	0	0	0
■□□	$(1-\beta)^2$	0	0
□■□	0	$(1-\beta)^2$	0
□□■	0	0	$(1-\beta)^2$
■■□	$\beta(1-\beta)$	$\beta(1-\beta)$	0
■□■	$\beta(1-\beta)$	0	$\beta(1-\beta)$
□■■	0	$\beta\,(1-\beta)$	$\beta(1-\beta)$
■■■	β^2	β^2	β^2

Index

University of BRISTOL

Universidad Carlos III de Madrid

Formulation

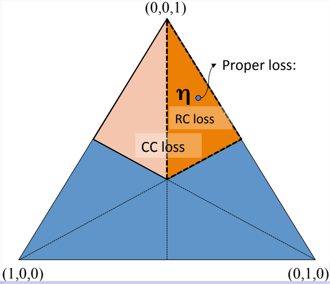


- *Conventional losses*
  - A function of the *true label* and the score: $\tilde{\Psi}(\mathbf{y}, \mathbf{f})$ in vector form, $\tilde{\boldsymbol{\Psi}}(\mathbf{f}) = \left(\tilde{\Psi}(\mathbf{e}^c_0, \mathbf{f}), \dots \tilde{\Psi}(\mathbf{e}^c_{c-1}, \mathbf{f})\right)$.
- *Weak losses*
  - A function of the *weak label* and the score: $\Psi(\mathbf{z}, \mathbf{f})$ in vector form, $\boldsymbol{\Psi}(\mathbf{f}) = (\Psi(\mathbf{b}_0, \mathbf{f}), \dots \Psi(\mathbf{b}_{d-1}, \mathbf{f}))$.
- We are interested in models based on Empirical Risk Minimization

$$\hat{R}_\Psi(\mathcal{S}) = \sum_{k=1}^{K} \Psi(\mathbf{z}_k, \mathbf{f}(\mathbf{x}_k)) \tag{1}$$

Types of loses

- Let $\mathbf{f}^*$ be a minimizer of $\mathbb{E}_{\mathbf{y}}\{\tilde{\Psi}(\mathbf{y}, \mathbf{f})\}$
- Three types of losses:
  1. *Proper*: a loss for posterior class probability estimation $\mathbf{f}^* = \boldsymbol{\eta}$
  2. *Ranking Calibrated (RC)*: A loss for ranking classes $f_i^* > f_j^* \Leftrightarrow \eta_i > \eta_j$
  3. *Classification Calibrated (CC)*: A loss to mininimize errors $f_i^* > \max_{j \neq i} f_j^* \Leftrightarrow \eta_i > \max_{j \neq i} \eta_j$



Weak loss

Theorem

Consider a weak loss Ψ *and a mixing matrix* $\mathbf{M}$*, and let the* equivalent *(conventional) loss* $\tilde{\Psi}$ *be given by*

$$\tilde{\boldsymbol{\Psi}}(\mathbf{f}) = \mathbf{M}^\intercal \boldsymbol{\Psi}(\mathbf{f}) \tag{2}$$

- Ψ *is (strictly)* $\mathbf{M}$*-proper iff* $\tilde{\Psi}$ *is (strictly) proper.*
- Ψ *is* $\mathbf{M}$*-RC iff* $\tilde{\Psi}$ *is RC.*
- Ψ *is* $\mathbf{M}$*-CC iff* $\tilde{\Psi}$ *is CC.*

Index

Index

Conventional loss into a weak loss

- We construct the weak loss in vector form as

$$\boldsymbol{\Psi}(\mathbf{f}) = \tilde{\mathbf{Y}}^\intercal \tilde{\boldsymbol{\Psi}}(\mathbf{f}) \tag{3}$$

  - Where $\tilde{\mathbf{Y}}$ is a weight matrix (we call it *virtual label matrix*)

- The weak loss for a weak label $\mathbf{b}_i$ can be written as

$$\Psi(\mathbf{f}, \mathbf{b}_i) = \tilde{\mathbf{y}}_i^\intercal \tilde{\boldsymbol{\Psi}}(\mathbf{f}) \tag{4}$$

  - where $\tilde{\mathbf{y}}_i$ is the i-th column of $\tilde{\mathbf{Y}}$

- For a conventional loss and a clean label $\tilde{\mathbf{y}}$

$$\tilde{\Psi}(\mathbf{f}, \mathbf{y}) = \mathbf{y}^\intercal \tilde{\boldsymbol{\Psi}}(\mathbf{f}) \tag{5}$$

- i-th column of $\tilde{\mathbf{Y}}$ is a *virtual label vector*

University of BRISTOL Universidad Carlos III de Madrid

Conventional loss into a weak loss

- We construct the weak loss in vector form as

$$\boldsymbol{\psi}(\mathbf{f}) = \tilde{\mathbf{Y}}^\intercal \tilde{\boldsymbol{\psi}}(\mathbf{f}) \tag{3}$$

  - Where $\tilde{\mathbf{Y}}$ is a weight matrix (we call it *virtual label matrix*)
- The weak loss for a weak label $\mathbf{b}_i$ can be written as

$$\psi(\mathbf{f}, \mathbf{b}_i) = \tilde{\mathbf{y}}_i^\intercal \tilde{\boldsymbol{\psi}}(\mathbf{f}) \tag{4}$$

  - where $\tilde{\mathbf{y}}_i$ is the i-th column of $\tilde{\mathbf{Y}}$
- For a conventional loss and a clean label $\tilde{\mathbf{y}}$

$$\tilde{\psi}(\mathbf{f}, \mathbf{y}) = \mathbf{y}^\intercal \tilde{\boldsymbol{\psi}}(\mathbf{f}) \tag{5}$$

- i-th column of $\tilde{\mathbf{Y}}$ is a *virtual label vector*

Conventional loss into a weak loss

- We construct the weak loss in vector form as

$$\boldsymbol{\Psi}(\mathbf{f}) = \tilde{\mathbf{Y}}^{\intercal}\tilde{\boldsymbol{\Psi}}(\mathbf{f}) \tag{3}$$

  - Where $\tilde{\mathbf{Y}}$ is a weight matrix (we call it *virtual label matrix*)
- The weak loss for a weak label $\mathbf{b}_i$ can be written as

$$\Psi(\mathbf{f}, \mathbf{b}_i) = \tilde{\mathbf{y}}_i^{\intercal}\tilde{\boldsymbol{\Psi}}(\mathbf{f}) \tag{4}$$

  - where $\tilde{\mathbf{y}}_i$ is the i-th column of $\tilde{\mathbf{Y}}$
- For a conventional loss and a clean label $\tilde{\mathbf{y}}$

$$\tilde{\Psi}(\mathbf{f}, \mathbf{y}) = \mathbf{y}^{\intercal}\tilde{\boldsymbol{\Psi}}(\mathbf{f}) \tag{5}$$

- i-th column of $\tilde{\mathbf{Y}}$ is a *virtual label vector*

Index

Mixing matrix is known

Theorem ([Cid-Sueiro et al., 2014])

Given a strictly proper loss $\tilde{\Psi}(\mathbf{f}, \mathbf{y})$ *and a virtual label matrix* $\tilde{\mathbf{Y}}$*, the weak loss* $\boldsymbol{\Psi}(\mathbf{f}) = \tilde{\mathbf{Y}}^\intercal \tilde{\boldsymbol{\Psi}}(\mathbf{f})$ *is strictly* $\mathcal{M}$*-proper, for any* $\mathbf{M} \in \mathcal{M}$ *such that* $\tilde{\mathbf{Y}}\mathbf{M} = \mathbf{I}$.

1. Compute a left inverse of the mixing matrix **M** that we call *virtual label matrix* $\tilde{\mathbf{Y}}$
2. *Substitute* each weak label $\mathbf{b_i}$ for its corresponding column $\tilde{\mathbf{y}}_\mathbf{i}$
3. Train with the conventional loss $\tilde{\Psi}$

Mixing matrix is unknown: if assume quasi_IPL

- The following $\mathbf{\Psi}$ loss is proper for any quasi-independent mixing matrix.

1. Take any label-based proper loss, e.g. the cross entropy:

$$\tilde{\mathbf{\Psi}}(\mathbf{f}) = \log(\mathbf{f}) \tag{6}$$

2. Take $\tilde{\mathbf{y}}$ given by

$$\tilde{y}_j = \left[\begin{array}{ll} 1 & z_j = 1 \\ -\frac{|\mathbf{z}|-1}{c-|\mathbf{z}|} & z_j = 0 \end{array} \right. \tag{7}$$

3. Take $\Psi(\mathbf{z}, \mathbf{f}) = \tilde{\mathbf{y}}^\intercal \tilde{\mathbf{\Psi}}(\mathbf{f})$

Mixing matrix is unknown: if assume IPL

- Mixing matrix M of the form IPL doesn't have proper weak loss
- M-RC and M-CC losses if we train with the weak labels

Implementation

- Common implementations of Cross-entropy *ignore* all except the true class
- They can not deal with *negative labels*
- We used *Brier Score* as is proper and does not have the previous problems
- We use the *Moore-Penrose* pseudoinverse $\tilde{\mathbf{Y}} = (\mathbf{M}^{T}\mathbf{M})^{-1}\mathbf{M}^{\mathsf{T}}$

Implementation

- Common implementations of Cross-entropy *ignore* all except the true class
- They can not deal with *negative labels*
- We used *Brier Score* as is proper and does not have the previous problems
- We use the *Moore-Penrose* pseudoinverse $\tilde{\mathbf{Y}} = (\mathbf{M}^T\mathbf{M})^{-1}\mathbf{M}^T$

Implementation

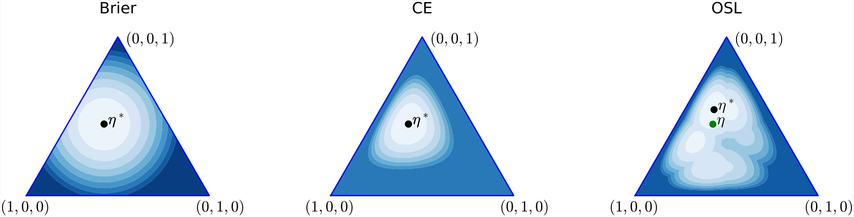
- Common implementations of Cross-entropy *ignore* all except the true class
- They can not deal with *negative labels*
- We used *Brier Score* as is proper and does not have the previous problems
- We use the *Moore-Penrose* pseudoinverse $\tilde{\mathbf{Y}} = (\mathbf{M}^T\mathbf{M})^{-1}\mathbf{M}^{\intercal}$

Implementation

- Common implementations of Cross-entropy *ignore* all except the true class
- They can not deal with *negative labels*
- We used *Brier Score* as is proper and does not have the previous problems
- We use the *Moore-Penrose* pseudoinverse $\tilde{\mathbf{Y}} = (\mathbf{M}^T\mathbf{M})^{-1}\mathbf{M}^\intercal$

Convexity

- if the loss $\tilde{\Psi}(\mathbf{f}, \tilde{\mathbf{y}}_i)$ is **M**-proper, its conditional expectation is a proper loss



- Other method: Optimistic Superset Learning

Index

Index

Datasets

- 31 real-world datasets
  - 3 to 20 classes
  - No more than 11,000 instances
  - All standardized
  - Available from openml.org
- Generating synthetic weak labels
  - 5 types of random mixing matrices
  - 4 levels of noise (increasing β)
  - Every combination repeated 10 times

Datasets

- 31 real-world datasets
  - 3 to 20 classes
  - No more than 11,000 instances
  - All standardized
  - Available from openml.org
- Generating synthetic weak labels
  - 5 types of random mixing matrices
  - 4 levels of noise (increasing β)
  - Every combination repeated 10 times

Models and learning method

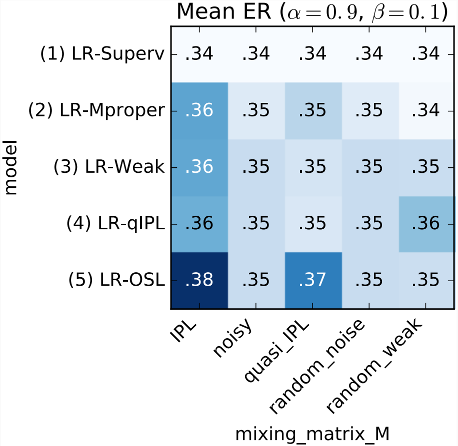
- Models
  - *LR*: Logistic Regression
  - *FNN*: Feed-Forward Neural Network
- Learning method
  - *Superv*: Using true labels
  - *Mproper*: Knowing **M**
  - *Weak*: Using weak labels (assuming IPL)
  - *quasiIPL*: Using virtual labels assuming quasi_IPL
  - *OSL*: Optimistic Superset Loss [Hüllermeier and Cheng, 2015]

Models and learning method

- Models
  - *LR*: Logistic Regression
  - *FNN*: Feed-Forward Neural Network
- Learning method
  - *Superv*: Using true labels
  - *Mproper*: Knowing **M**
  - *Weak*: Using weak labels (assuming IPL)
  - *quasiIPL*: Using virtual labels assuming quasi_IPL
  - *OSL*: Optimistic Superset Loss [Hüllermeier and Cheng, 2015]

Index

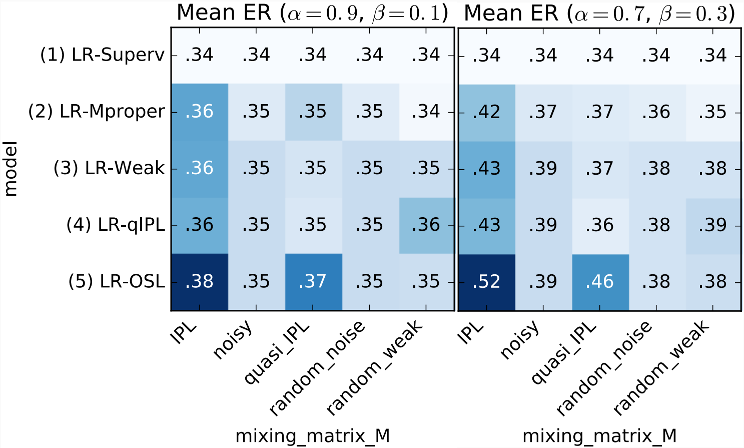
Error rate Logistic Regression



Mean ER ($\alpha = 0.9$, $\beta = 0.1$)

model \ mixing_matrix_M	IPL	noisy	quasi_IPL	random_noise	random_weak
(1) LR-Superv	.34	.34	.34	.34	.34
(2) LR-Mproper	.36	.35	.35	.35	.34
(3) LR-Weak	.36	.35	.35	.35	.35
(4) LR-qIPL	.36	.35	.35	.35	.36
(5) LR-OSL	.38	.35	.37	.35	.35

Error rate Logistic Regression



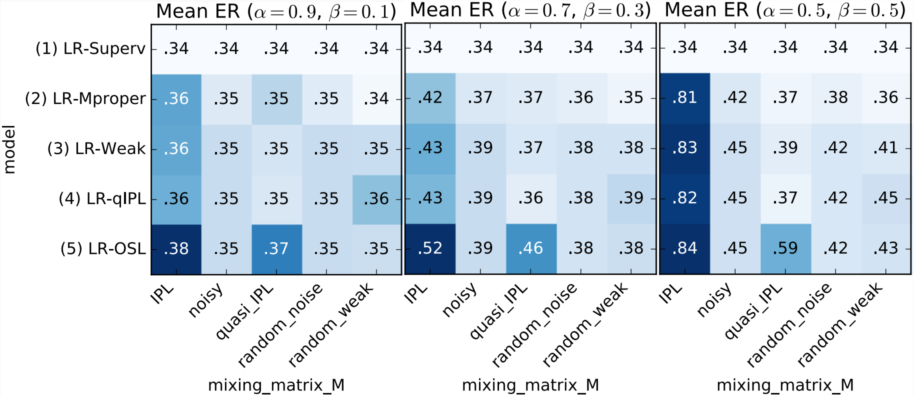
Mean ER ($\alpha = 0.9$, $\beta = 0.1$)

model	IPL	noisy	quasi_IPL	random_noise	random_weak
(1) LR-Superv	.34	.34	.34	.34	.34
(2) LR-Mproper	.36	.35	.35	.35	.34
(3) LR-Weak	.36	.35	.35	.35	.35
(4) LR-qIPL	.36	.35	.35	.35	.36
(5) LR-OSL	.38	.35	.37	.35	.35

Mean ER ($\alpha = 0.7$, $\beta = 0.3$)

model	IPL	noisy	quasi_IPL	random_noise	random_weak
(1) LR-Superv	.34	.34	.34	.34	.34
(2) LR-Mproper	.42	.37	.37	.36	.35
(3) LR-Weak	.43	.39	.37	.38	.38
(4) LR-qIPL	.43	.39	.36	.38	.39
(5) LR-OSL	.52	.39	.46	.38	.38

Error rate Logistic Regression



Mean ER ($\alpha = 0.9$, $\beta = 0.1$)

model \ mixing_matrix_M	IPL	noisy	quasi_IPL	random_noise	random_weak
(1) LR-Superv	.34	.34	.34	.34	.34
(2) LR-Mproper	.36	.35	.35	.35	.34
(3) LR-Weak	.36	.35	.35	.35	.35
(4) LR-qIPL	.36	.35	.35	.35	.36
(5) LR-OSL	.38	.35	.37	.35	.35

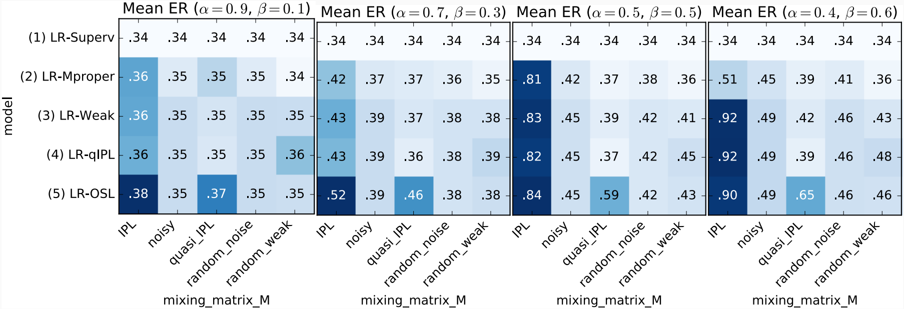
Mean ER ($\alpha = 0.7$, $\beta = 0.3$)

model \ mixing_matrix_M	IPL	noisy	quasi_IPL	random_noise	random_weak
(1) LR-Superv	.34	.34	.34	.34	.34
(2) LR-Mproper	.42	.37	.37	.36	.35
(3) LR-Weak	.43	.39	.37	.38	.38
(4) LR-qIPL	.43	.39	.36	.38	.39
(5) LR-OSL	.52	.39	.46	.38	.38

Mean ER ($\alpha = 0.5$, $\beta = 0.5$)

model \ mixing_matrix_M	IPL	noisy	quasi_IPL	random_noise	random_weak
(1) LR-Superv	.34	.34	.34	.34	.34
(2) LR-Mproper	.81	.42	.37	.38	.36
(3) LR-Weak	.83	.45	.39	.42	.41
(4) LR-qIPL	.82	.45	.37	.42	.45
(5) LR-OSL	.84	.45	.59	.42	.43

Error rate Logistic Regression



Mean ER ($\alpha=0.9$, $\beta=0.1$)

model	IPL	noisy	quasi_IPL	random_noise	random_weak
(1) LR-Superv	.34	.34	.34	.34	.34
(2) LR-Mproper	.36	.35	.35	.35	.34
(3) LR-Weak	.36	.35	.35	.35	.35
(4) LR-qIPL	.36	.35	.35	.35	.36
(5) LR-OSL	.38	.35	.37	.35	.35

Mean ER ($\alpha=0.7$, $\beta=0.3$)

model	IPL	noisy	quasi_IPL	random_noise	random_weak
(1) LR-Superv	.34	.34	.34	.34	.34
(2) LR-Mproper	.42	.37	.37	.36	.35
(3) LR-Weak	.43	.39	.37	.38	.38
(4) LR-qIPL	.43	.39	.36	.38	.39
(5) LR-OSL	.52	.39	.46	.38	.38

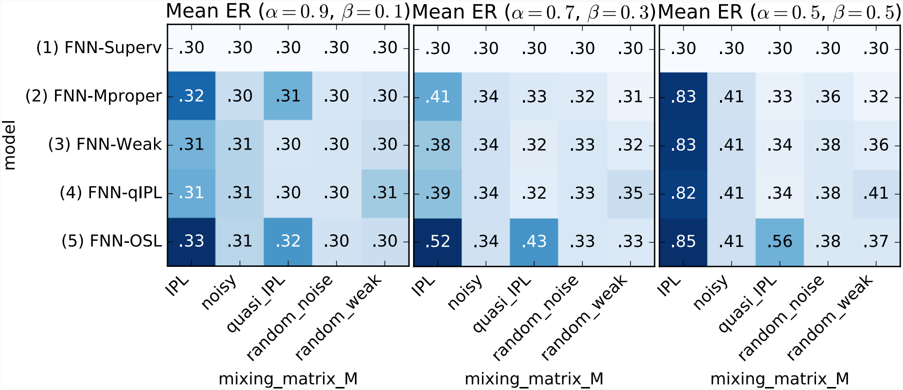
Mean ER ($\alpha=0.5$, $\beta=0.5$)

model	IPL	noisy	quasi_IPL	random_noise	random_weak
(1) LR-Superv	.34	.34	.34	.34	.34
(2) LR-Mproper	.81	.42	.37	.38	.36
(3) LR-Weak	.83	.45	.39	.42	.41
(4) LR-qIPL	.82	.45	.37	.42	.45
(5) LR-OSL	.84	.45	.59	.42	.43

Mean ER ($\alpha=0.4$, $\beta=0.6$)

model	IPL	noisy	quasi_IPL	random_noise	random_weak
(1) LR-Superv	.34	.34	.34	.34	.34
(2) LR-Mproper	.51	.45	.39	.41	.36
(3) LR-Weak	.92	.49	.42	.46	.43
(4) LR-qIPL	.92	.49	.39	.46	.48
(5) LR-OSL	.90	.49	.65	.46	.46

Error rate Feed-Forward Neural Network



Mean ER ($\alpha = 0.9$, $\beta = 0.1$)

model	IPL	noisy	quasi_IPL	random_noise	random_weak
(1) FNN-Superv	.30	.30	.30	.30	.30
(2) FNN-Mproper	.32	.30	.31	.30	.30
(3) FNN-Weak	.31	.31	.30	.30	.30
(4) FNN-qIPL	.31	.31	.30	.30	.31
(5) FNN-OSL	.33	.31	.32	.30	.30

Mean ER ($\alpha = 0.7$, $\beta = 0.3$)

model	IPL	noisy	quasi_IPL	random_noise	random_weak
(1) FNN-Superv	.30	.30	.30	.30	.30
(2) FNN-Mproper	.41	.34	.33	.32	.31
(3) FNN-Weak	.38	.34	.32	.33	.32
(4) FNN-qIPL	.39	.34	.32	.33	.35
(5) FNN-OSL	.52	.34	.43	.33	.33

Mean ER ($\alpha = 0.5$, $\beta = 0.5$)

model	IPL	noisy	quasi_IPL	random_noise	random_weak
(1) FNN-Superv	.30	.30	.30	.30	.30
(2) FNN-Mproper	.83	.41	.33	.36	.32
(3) FNN-Weak	.83	.41	.34	.38	.36
(4) FNN-qIPL	.82	.41	.34	.38	.41
(5) FNN-OSL	.85	.41	.56	.38	.37

Index

Conclusion

- *Simple method* just changing the weak labels into virtual labels
- Adapts *existing classifiers* based on empirical minimization of losses
- Possible to add priors to improve results
- Best results if we can estimate the mixing matrix **M**

Future work

- Real datasets with weak labels
- Merging datasets containing different types of noise
- Different noise depending on the input space

Cid-Sueiro, J., García-García, D., and Santos-Rodríguez, R. (2014).
Consistency of losses for learning from weak labels.
In *European Conference on Machine Learning and Knowledge Discovery in Databases*, pages 197–210. Springer Berlin Heidelberg.

Hüllermeier, E. and Cheng, W. (2015).
Superset learning based on generalized loss minimization.
In *European Conference on Machine Learning and Knowledge Discovery in Databases*, pages 260–275.

Adapting supervised classification algorithms to weak label scenarios

Miquel Perelló-Nieto[1], Raúl Santos-Rodríguez[1], and Jesús Cid-Sueiro[2]

[1]University of Bristol, UK
[2]Universidad Carlos III de Madrid, Spain

Email: [1]{Miquel.PerelloNieto, enrsr}@bristol.ac.uk, [2]jcid@tsc.uc3m.es

October 26, 2017