Perelló-Nieto, M., Santos-Rodríguez, R., & Cid-Sueiro, J. (2017). Adapting Supervised Classification Algorithms to Arbitrary Weak Label Scenarios. In *Advances in Intelligent Data Analysis XVI: 16th International Symposium, IDA 2017, London, UK, October 26–28, 2017, Proceedings* (pp. 247-259). (Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics); Vol. 10584 LNCS). Springer Verlag. https://doi.org/10.1007/978-3-319-68765-0_21

Adapting supervised classification algorithms to arbitrary weak label scenarios

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

[1] University of Bristol, UK
{miquel.perellonieto,enrsr}@bristol.ac.uk
[2] Universidad Carlos III de Madrid, Spain
jcid@tsc.uc3m.es

Abstract. In many real-world problems, labels are often weak, meaning that each instance is labelled as belonging to one of several candidate categories, at most one of them being true. Recent theoretical contributions have shown that it is possible to construct proper losses or classification calibrated losses for weakly labelled classification scenarios by means of a linear transformation of conventional proper or classification calibrated losses, respectively. However, how to translate these theoretical results into practice has not been explored yet. This paper discusses both the algorithmic design and the potential advantages of this approach, analyzing consistency and convexity issues arising in practical settings, and evaluating the behavior of such transformations under different types of weak labels.

Keywords: Weak labels, noisy labels, proper losses

1 Introduction

Most machine learning algorithms are grounded on two common assumptions: (1) a pool of annotated examples is available for training, and (2) the labelling process satisfies some nice statistical properties, e.g., balanced label proportions or statistical independence. However, in practice, real datasets pose major challenges regarding the quality of the labels, from label noise to partial supervision.

In the last decade several authors addressed these and similar tasks using different terminologies depending on the specific properties and assumptions of the scenarios at hand. For instance, learning from *partial labels* [6, 11, 5], *multiple labels* [10] and *ambiguous labels* [8], describe settings where each sample is labeled using a subset of classes, required to contain the true label. On the other hand, *crowd learning* [13, 12] assumes that we may have access to multiple labels provided by a number of different annotators, that are bound to sometimes disagree when labelling the same example. A related problem, *noisy labels* [2, 14],

JCS is supported by the TEC2014-52289-R project funded by the Spanish MEC. MPN and RSR are supported by the SPHERE IRC funded by the UK Engineering and Physical Sciences Research Council (EP/K031910/1).

restricts the setting to a unique label per sample, but in this case labels have some constant probability of being flipped. Finally, *superset learning* [9] generalizes many of the previous approaches and allows for each example to be linked to a subset of classes that contains the true outcome but may also encompass additional ones. In this paper, our interpretation of the weak label paradigm also accounts for each instance being labelled as belonging to one of several candidate categories, at most one of them being true, but we do not require the true label to be present.

In [3] we suggested a general procedure to transform a standard (i.e. fully-supervised) proper loss into a weak loss that is also proper, in the sense that posterior class probabilities can be estimated provided that the label mixing process is restricted to lie in certain linear subspace. Recently, in [4] we analyzed the conditions under which the true class can be inferred from weak labels. In this paper we built upon this previous theoretical results to describe a simple algorithmic approach to seamlessly adapt existing classification algorithms that are based on the empirical minimization of proper or classification calibrated losses, in such a way that they explicitly take into account the mixing process underlying the generation of the annotations. In short, the contributions of this work are twofold:

- We depict a transparent procedure for the machine learning practitioner to transform weak labels into what we refer to as *virtual labels*, that can then be used within standard out-of-the-shelf machine learning toolboxes, providing advice on practical implementation issues.
- We thoroughly test the approach studying realistic scenarios in which, (1) partial information might be available regarding the relationship between some true and weak labels and, (2) no information at all is revealed.

The remainder of the paper is organized as follows: the problem of learning from weak labels is formulated in Sec. 2. Some results on losses for weak labels are reviewed in Sec. 3, and the algorithmic design is detailed in Sec. 4. In Sec. 5 we analyse five common weakly supervised case studies. Finally, we state some conclusions in Sec. 6.

2 Formulation

2.1 Notation

Vectors are written in boldface, matrices in boldface capital and sets in calligraphic letters. For any integer n, $\mathbf{e}_i^n$ is a n-dimensional unit vector with all zero components apart from the i-th component which is equal to one, and $\mathbb{1}_n$ is a n-dimensional all-ones vector. Superindex $^\intercal$ denotes transposition. We will use $\Psi()$ to denote a loss based on weak labels (for brevity, "weak loss"), and $\tilde{\Psi}$ to losses based on the true class. The number of classes is c, and the number of possible weak label vectors is $d \leq 2^c$. $|\mathbf{z}|$ is the number of nonzero elements in $\mathbf{z}$. The set of all $d \times c$ matrices with stochastic columns is

$\mathcal{M} = \{\mathbf{M} \in [0,1]^{d\times c} : \mathbf{M}^\intercal \mathbb{1}_d = \mathbb{1}_c\}$, and the simplex of n-dimensional probability vectors is $\mathcal{P}_n = \{\mathbf{p} \in [0,1]^n : \sum_{i=0}^{n-1} p_i = 1\}$.

2.2 Learning from weak labels

Let $\mathcal{X}$ be a sample space, $\mathcal{Y} = \{\mathbf{e}_j^c, j = 0, 1, \ldots, c-1\}$ a set of labels, and $\mathcal{Z} = \{\mathbf{b}_1, \ldots, \mathbf{b}_d\} \subset \{0,1\}^c$ a set of weak or partial label vectors. Sample $(\mathbf{x}, \mathbf{z}) \in \mathcal{X} \times \mathcal{Z}$ is drawn from an unknown distribution P.

Weak label vector $\mathbf{z} \in \mathcal{Z}$ is a noisy version of the actual true class $\mathbf{y} \in \mathcal{Y}$. A common assumption [5, 10, 1, 7] is that the true class is always present in $\mathbf{z}$, i.e., $z_j = 1$ when $y_j = 1$, but this assumption is not required in our setting. We assume that $\mathcal{Z}$ contains only weak labels with nonzero probability (i.e. $P\{\mathbf{z} = \mathbf{b}\} > 0$ for any $\mathbf{b} \in \mathcal{Z}$). The dependency between $\mathbf{z}$ and $\mathbf{y}$ is modelled through an arbitrary $d \times c$ conditional mixing probability matrix $\mathbf{M}(\mathbf{x}) \in \mathcal{M}$ with components

$$m_{ij}(\mathbf{x}) = P\{\mathbf{z} = \mathbf{b}_i | y_j = 1, \mathbf{x}\} \tag{1}$$

where $\mathbf{b}_i \in \mathcal{Z}$ is the i-th element of $\mathcal{Z}$. Defining posterior probability vectors $\mathbf{p}(\mathbf{x})$ and $\boldsymbol{\eta}(\mathbf{x})$ with components $p_i = P\{\mathbf{z} = \mathbf{b}_i | \mathbf{x}\}$ and $\eta_j = P\{\mathbf{y} = \mathbf{e}_j^c | \mathbf{x}\}$, we can write $\mathbf{p}(\mathbf{x}) = \mathbf{M}(\mathbf{x})\boldsymbol{\eta}(\mathbf{x})$. In general, the dependency with $\mathbf{x}$ will be omitted and we will write, for instance, $\mathbf{p} = \mathbf{M}\boldsymbol{\eta}$. The mixing matrix could depend on $\mathbf{x}$, though a constant mixing matrix is a common assumption [13, 10, 1, 7], as well as the statistical independence of the incorrect labels [10, 1, 7]. Assuming a constant matrix is not required in our analysis. Any property derived for $\mathbf{M}$ can be extended to a property that must be satisfied by $\mathbf{M}(\mathbf{x})$ for all $\mathbf{x}$.

The goal is to infer $\mathbf{y}$ given $\mathbf{x}$ without knowing model P. To do so, a set of i.i.d. weakly labelled samples, $\mathcal{S} = \{(\mathbf{x}_k, \mathbf{z}_k), k = 1, \ldots, K\} \sim P$ is available. True classes $\mathbf{y}_k$ are not observed. In this paper we are interested in algorithms based on the minimization of an empirical risk

$$\hat{R}_\Psi(\mathcal{S}) = \sum_{k=1}^{K} \Psi(\mathbf{z}_k, \mathbf{f}(\mathbf{x}_k)) \tag{2}$$

where $\Psi(\mathbf{z}, \mathbf{f})$ is a weak loss function that takes a weak label (instead of the true label) as an argument, and $\mathbf{f}(\mathbf{x})$ is a scoring function. The class prediction is computed through some function $\text{pred}(\mathbf{x}) \in \text{argmax}_i\{f_i(\mathbf{x})\}$.

3 Transforming a conventional loss into a weak loss

This section summarizes some of the theoretical results in [3] and [4] that have practical implications on the design of algorithms for weakly labeled datasets.

3.1 Virtual labels

We will consider weak loss functions that can be computed as linear combinations of conventional loss functions for clean labels. Defining the vector representation

of the weak loss as $\boldsymbol{\Psi}(\mathbf{f}) = (\Psi(\mathbf{b}_0, \mathbf{f}), \dots \Psi(\mathbf{b}_{d-1}, \mathbf{f}))$, we construct losses following

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

where $\tilde{\boldsymbol{\Psi}}$ is a vector representation of a conventional loss $\tilde{\Psi}(\mathbf{y}, \mathbf{f})$, that is, $\tilde{\boldsymbol{\Psi}}(\mathbf{f}) = \left(\tilde{\Psi}(\mathbf{e}_0^c, \mathbf{f}), \dots \tilde{\Psi}(\mathbf{e}_{c-1}^c, \mathbf{f})\right)$ and $\tilde{\mathbf{Y}}$ is a weight matrix.

Note that 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 $\tilde{\mathbf{Y}}$. By comparing this expression with the loss for a clean label $\mathbf{y}$

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

we can interpret the i-th column of $\tilde{\mathbf{Y}}$ as a *virtual* label vector. The weak loss can thus be computed by replacing the true label by the virtual label corresponding to the observed weak label. For this reason we will call $\tilde{\mathbf{Y}}$ a *virtual label matrix*.

3.2 Properness and classification calibration

Linear transformations in the form (3) have ben studied in [4]. In particular, the authors studied the conditions on the virtual matrix that guarantee that the weak loss is **M**-proper or **M**-classification calibrated.

A loss function is said to be **M**-proper if the minimizer of the expected loss is the true posterior class probability, i.e. $\boldsymbol{\eta} \in \arg\min_{\mathbf{f}} \mathbb{E}_{\mathbf{z}}\{\Psi(\mathbf{z}, \mathbf{f})\}$, where $\boldsymbol{\eta}$ is the probability vector with components $\eta_j = P\{y_j = 1\}$. The loss is strictly proper if $\boldsymbol{\eta}$ is the unique minimizer. If posterior class probability estimates are not required and the main goal is to minimize classification error, classification calibration can be enough. We say that a weak loss is **M**-classification calibrated (or **M**-CC) if $\mathbf{f}^* \in \arg\min_{\mathbf{f}} \mathbb{E}_{\mathbf{z}}\{\Psi(\mathbf{z}, \mathbf{f})\}$ satisfies $(\eta_i > \max_{j \neq c} \eta_j \Rightarrow f_i^* > \max_{j \neq c} f_j^*)$, that is, a class with maximum a posteriori probability is also a class with the highest score value in $\mathbf{f}$.

Example The difference between a proper and a classification calibrated loss is illustrated in Fig. 1, which represents the probability simplex for a 3-class problem. Every point in the triangle is a probability vector $\mathbf{p} = (p_0, p_1, p_2)$. If the true posterior class probability for a given input $\mathbf{x}$ is given by point $\boldsymbol{\eta}$, the minimizer of the risk associated with a proper loss should be exactly $\boldsymbol{\eta}$. Classification calibration is less restrictive as the minimizer of the risk associated to a CC loss can be any point in the lighter region around $(0, 0, 1)$. Intuitively, the choice of a good virtual matrix for a given mixing matrix **M** is less restrictive for an **M**-CC loss than for an **M**-proper loss. This is formalized below.

3.3 Constructing a weak loss when the mixing matrix is known

If **M** is know, we might be interested in systematic procedures to design an **M**-proper or **M**-CC loss. The main result is summarized in the following theorems.

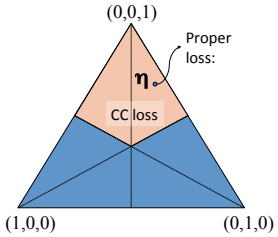

Fig. 1: Location of the minimizers of a proper loss (a single point) or a classification calibrated loss (the region around $\boldsymbol{\eta}$, for a true probability $\boldsymbol{\eta}$

Theorem 1 ([4]). *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 $\tilde{\mathbf{Y}}\mathbf{M} = \mathbf{I}$.*

The theorem states that any left inverse of the mixing matrix $\mathbf{M}$ is a valid virtual label matrix. The analogous result for CC losses is the following.

Theorem 2 ([4]). *Given a CC loss $\tilde{\Psi}(\mathbf{f}, \mathbf{y})$ and a matrix $\tilde{\mathbf{Y}}$, the weak loss $\boldsymbol{\Psi}(\mathbf{f}) = \tilde{\mathbf{Y}}^\intercal \tilde{\boldsymbol{\Psi}}(\mathbf{f})$ is $\mathbf{M}$-classification calibrated, for any $\mathbf{M}$ such that $\tilde{\mathbf{Y}}\mathbf{M} = \lambda\mathbf{I} + \mathbb{1}_c\mathbf{v}^\intercal$, for some $\lambda \in \mathbb{R}^+$ and $\mathbf{v} \in \mathbb{R}^c$.*

As a consequence, any matrix in the form $\tilde{\mathbf{Y}} = (\lambda\mathbf{I} + \mathbb{1}_c\mathbf{v}^\intercal)\tilde{\mathbf{Y}}_0$, where $\tilde{\mathbf{Y}}_0$ is an arbitrary left inverse of the mixing matrix, is a valid virtual label matrix to construct a classification-calibrated loss. Note that, parameters λ and $\mathbf{v}$ provide more degrees of freedom to construct a CC loss with respect to a proper loss.

3.4 Constructing a weak loss when the mixing matrix is unknown

If the mixing matrix is unknown, the classification problem becomes unresolvable, as the mixing matrix can not be inferred from a weakly labelled dataset. For instance, without any additional side information, $\mathbf{M} = \mathbf{I}$ is not distinguishable from any random permutation of its rows or columns. However, theorems 1 and 2 show that a weak loss $\boldsymbol{\Psi}(\mathbf{f}) = \tilde{\mathbf{Y}}^\intercal \tilde{\boldsymbol{\Psi}}(\mathbf{f})$ can be proper for a large set of different mixing matrices. Under some conditions, loss functions can be constructed to be $\mathbf{M}$-proper or $\mathbf{M}$-cc over pre-specified sets of mixing matrices. In particular, we can derive generic proper losses that are admissible under some general independence assumptions on the mixing process.

Case 1: Losses for quasi independent labels Consider the conditional probability model given by

$$P(\mathbf{z}|\mathbf{y} = \mathbf{e}_m^c) = \begin{bmatrix} z_m \beta_{m,|\mathbf{z}|} & |\mathbf{z}| < c \\ 0 & |\mathbf{z}| = c \text{ or } |\mathbf{z}| = 0 \end{bmatrix} \tag{6}$$

where coefficients $\beta_{m,n}$ satisfy the linear constraint

$$\sum_{n=1}^{c} \binom{c-1}{n-1} \beta_{m,n} = 1 \tag{7}$$

In [4], it is shown that, for some particular values of coefficient $\beta_{m,|z|}$, this model is almost equivalent to the case where the observation of a class in the weak label vector does not depend on all other classes, but only on the true class (the model would be equivalent in the probability of observing all classes would be nonzero). However, it can be shown that the virtual label matrix $\tilde{\mathbf{Y}}$ with virtual label vectors

$$\tilde{y}_j = \begin{bmatrix} 1 & z_j = 1 \\ -\frac{|\mathbf{z}|-1}{c-|\mathbf{z}|} & z_j = 0 \end{bmatrix} \tag{8}$$

(the case $|\mathbf{z}| = c$ is ignored), is admissible for the quasi independent model, no matter what the specific values of parameter $\beta_{m,n}$ are.

Case 2: Classification calibrated losses for independent labels Another interesting choice for the virtual label matrix is to take $\tilde{\mathbf{y}}_i = \mathbf{b}_i$, i.e. taking the virtual label vectors equal to the weak label vectors. It can be shown that, though a loss based on this matrix is not **M**-proper, it is **M**-CC for the independent label model

$$P(\mathbf{z}|\mathbf{y} = \mathbf{e}_m^c) = \alpha^{z_m}(1-\alpha)^{1-z_m}\beta^{|\mathbf{z}|-1}(1-\beta)^{c-|\mathbf{z}|} \tag{9}$$

4 Algorithmic design

Building upon the results in the previous Section, we now discuss the implementation of specific algorithms for weak labels, analyzing consistency and convexity issues arising in practical settings.

4.1 Replacing true labels with virtual labels

The analogy between eqs. (4) and (5) suggests that we can easily transform any conventional algorithm for clean labels into an algorithm minimizing a weak loss, by simply replacing true label vectors by the virtual label vectors corresponding to the observed weak labels. In practice, this is not always the case. Some specific implementations of algorithms for the minimization of some losses for clean labels do not work when the target vector has not the conventional form of all zeros but a single one.

For instance, assume the logistic regression model, where the scoring function is given by

$$f_i = \frac{\exp(\mathbf{w}_i^\intercal \mathbf{x})}{\sum_{k=0}^{c-1} \exp(\mathbf{w}_j^\intercal \mathbf{x})} \tag{10}$$

The cross entropy is a common choice of a proper loss for this model, and is given by $\tilde{\boldsymbol{\Psi}}(\mathbf{f}) = -\log(\mathbf{f})$ (where the log is the natural logarithm and it is computed component-wise). The gradient of the weak loss (4) with respect to the model weights is given by

$$g(\mathbf{w}_0, \ldots, \mathbf{w}_{c-1}) = \mathbf{x}^\intercal(b_i \cdot \mathbf{f} - \tilde{\mathbf{y}}_i) \tag{11}$$

where the coefficient $b_i = \mathbb{1}^\intercal \tilde{\mathbf{y}}_i$ is equal to one in a clean label case, so it is ignored in usual implementations of gradient based learning rules for the cross entropy. It is, however, required for learning from weak losses.

4.2 Consistency

When the loss function $\tilde{\boldsymbol{\Psi}}$ is not upper bounded, the implementation of an $\mathbf{M}$ proper loss may present consistency issues. For instance, for a logistic model with a cross entropy loss function, we have

$$\tilde{\Psi}(\mathbf{f}, \tilde{\mathbf{y}}_i) = \tilde{\mathbf{y}}_i^\intercal \mathbf{W}\mathbf{x} - (\mathbb{1}^\intercal \tilde{\mathbf{y}}_i) \log \left(\sum_{j=0}^{c-1} \exp(\mathbf{w}_j^\intercal \mathbf{x}) \right) \tag{12}$$

If the virtual labels contain negative components, it is not difficult to show that, for some values of $\mathbf{x}$, the loss is not bounded below and, for some datasets, the minimum empirical risk is $-\infty$. Thus, the minimizer of the empirical risk may not converge to the true posterior class probabilities.

This problem can be resolved by taking into account that $|\tilde{\Psi}(\mathbf{f}, \tilde{\mathbf{y}}_i)| \geq -\lambda \|\mathbf{W}\| - \max_i \|\tilde{\mathbf{y}}_i\| \log(C)$, where $\lambda = 2 \max_i \|\tilde{\mathbf{y}}_i\| \max_{\mathbf{x} \in \mathcal{S}} \|\mathbf{x}\|$. Thus, the regularized loss $\tilde{\Psi}_\lambda(\mathbf{f}, \tilde{\mathbf{y}}_i) = \tilde{\Psi}_\lambda(\mathbf{f}, \tilde{\mathbf{y}}_i) + \lambda \|\mathbf{W}\|$ is bounded below. In order to avoid these inconsistency issues, we have used the Brier score (square loss) $\tilde{\Psi}_\lambda(\mathbf{f}, \tilde{\mathbf{y}}_i) = \|\mathbf{y}_i - \mathbf{f}\|^2$, which is known to be proper, in our experiments.

4.3 Convexity

It can be show that, if the loss $\tilde{\Psi}(\mathbf{f}, \tilde{\mathbf{y}}_i)$ is $\mathbf{M}$-proper, its conditional expectation is a proper loss, which is necessarily a convex function of the true posterior, $\boldsymbol{\eta}$. This is illustrated in Fig. 2, which shows the contour plots of the expected loss for the Brier score (left) and Cross Entropy (center) for a given mixing matrix and a true posterior $\boldsymbol{\eta} = (0.35, 0.2, 0.45)$ on the 3 class probability simplex. The right plot shows the conditional expected value of the Optimistic Superset Loss [9], which consists on replacing the weak label by a tentative single-class label, selected as the class with the highest score among the candidate labels. Note that this loss is not convex and may have several local minima.

Note that, despite the conditional loss being a convex function of the score, the weak loss might be a non convex function of the model parameters, depending on the type of virtual label matrix. Consider, for instance, a parametric score function $\mathbf{f}_\mathbf{w}(\mathbf{x})$ and assume that the conventional loss $\tilde{\Psi}(\mathbf{f}_\mathbf{w}(\mathbf{x}), \mathbf{y})$ is convex on the model parameters $\mathbf{w}$, for any target vector $\mathbf{y} \in \mathcal{Y}$ (as, for instance, in logistic regression). If the virtual label contains negative values, then the weak loss $\tilde{\mathbf{Y}}\boldsymbol{\Psi}$ is in general not convex. This is a major difficulty in order to preserve convexity and properness as, if $\tilde{\mathbf{Y}}$ is a left inverse of a nonnegative probability matrix, it usually contains negative components. Fortunately, we can always construct a CC-loss with the appropriate selection of the virtual label matrix.

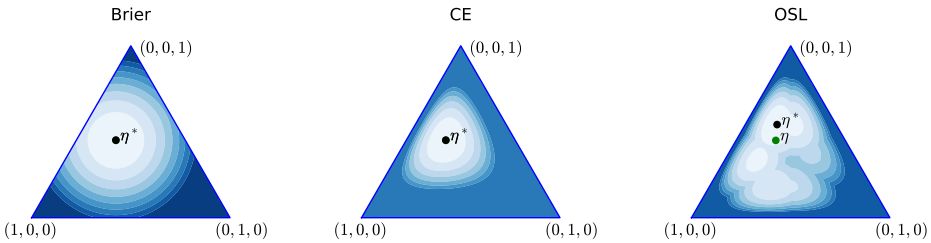

Fig. 2: Conditional expected loss over the 3 class probability simplex for a true posterior $\boldsymbol{\eta} = (0.35, 0.2, 0.45)$, for three different losses: Brier score (left), Cross Entropy (center) and Optimistic Superset Loss (right). $\boldsymbol{\eta}$ is the true posterior, and $\boldsymbol{\eta}^*$ is the minimizer. For the Brier score and the CE score, as expected, $\boldsymbol{\eta}$ and $\boldsymbol{\eta}^*$ coincide.

Theorem 3. *If $\tilde{\Psi}(\mathbf{f}_{\mathbf{w}}(\mathbf{x}), \mathbf{y})$ is a convex function of $\mathbf{w}$ for any $\mathbf{y}$ and $\tilde{\mathbf{Y}}\tilde{\boldsymbol{\Psi}}(\mathbf{f})$ is $\mathbf{M}$-proper, then the weak loss given by $\boldsymbol{\Psi}(\mathbf{f}) = \tilde{\mathbf{Y}}'\tilde{\boldsymbol{\Psi}}(\mathbf{f})$ is $\mathbf{M}$-CC and convex in $\mathbf{w}$, where*

$$\tilde{\mathbf{Y}}' = \tilde{\mathbf{Y}} - \mathbf{m}\mathbb{1}_d^\intercal \tag{13}$$

where $\mathbf{m}$ is the row-wise minimum of matrix $\mathbf{M}$

The proof is a direct consequence of: (1) $\tilde{\mathbf{Y}}'$ satisfies the conditions in Th. 3, (so the weak loss is CC), and (2) $\tilde{\mathbf{Y}}'$ has nonnegative components, so the weak loss is a conic combination of convex functions, so it is convex.

4.4 Selection of the virtual matrix

Note that, in general (if the rank of $\mathbf{M}$ is higher than c) the set of admissible virtual label matrices for a given mixing matrix is infinite. However, this does not mean that the choice of $\tilde{\mathbf{Y}}$ in this set is irrelevant. Different virtual label matrices show different performances. Though we have no theoretical evidence, we have experimentally found that a good choice for $\mathbf{Y}$ is the Moore-Penrose pseudoinverse $\tilde{\mathbf{Y}} = (\mathbf{M}^T\mathbf{M})^{-1}\mathbf{M}^\intercal$ (we assume $\mathbf{M}$ is not rank-deficient, otherwise the classification problem would be severely degenerated by the mixing process).

5 Experiments

In this section, we demonstrate the use of our framework on real-world data. We show how the proposed method can effectively transform existing classification algorithms so that they incorporate the available information regarding the label quality, thereby providing robust solutions to weakly supervised data that is not addressed when using out-of-the-box machine learning techniques.

Datasets We tested the models on 31 real-world datasets from `openml.org`. From the available datasets, we chose those with number of classes between 3 and 20, less than 11,000 instances and no missing values. Then, we sorted all the filtered datasets by highest impact and manually selected the final subset. Before training, all the categorical features were transformed into binary features using a one-hot encoding. Finally, every feature was standardised with mean zero and standard deviation one. Table 1 contains a summary of the datasets.

Table 1: Summary of the 31 datasets used in the experiments

name	size	features	classes	name	size	features	classes
GesturePhaseSegme.	9873	32	5	flags	194	120	8
JapaneseVowels	9961	14	9	glass	214	9	7
abalone	4177	10	3	iris	150	4	3
analcatdata_dmft	797	21	6	mfeat-zernike	2000	47	10
autoUniv-au6-1000	1000	44	8	page-blocks	5473	10	5
autoUniv-au6-750	750	44	8	pendigits	10992	16	10
autoUniv-au7-1100	1100	18	5	prnn_fglass	214	9	6
autoUniv-au7-500	500	18	5	satimage	6430	36	6
balance-scale	625	4	3	segment	2310	19	7
car	1728	21	4	vehicle	846	18	4
cardiotocography	2126	35	10	visualizing_liv.	130	27	5
collins	500	35	15	vowel	990	27	11
confidence	72	3	6	wine	178	13	3
diggle_table_a2	310	8	9	yeast	1484	8	10
ecoli	336	7	8	zoo	101	31	7
fl2000	67	16	5				

As all datasets have only one true label per sample, we needed to artificially weaken the labels. We simulated five common scenarios by generating different random sets of mixing matrices $\mathbf{M}$:

1. **Noisy**: Each sample has only one label. The weak label is the true label with probability α. The rest of the classes are equally probable. This represents scenarios where, in order to reduce annotation costs, the labels are bound to contain mistakes.
2. **Random_noise**: Each sample has only one label. The weak label is the true label with at least probability α. The rest of the classes have probability β_m, initially drawn from a uniform distribution. The value of α defines the degree of supervision, from fully supervised to a completely random mixing matrix $\mathbf{M}$.
3. **IPL**: Each sample can have multiple labels. In this case the true label has a probability α of appearing, while other labels are present with a certain probability β. This scenario occurs in complex classification tasks where multiple annotators label the samples but there is not known ground truth.

4. **Quasi_IPL**: Each sample can have multiple labels. In this case the true label is always present but other labels may appear with certain probability β. This scenario could be applied to the identification of an animal subspecies given that we know its taxonomic parents.
5. **Random_weak**: Each sample can have multiple labels. In this case all the possible weak labels may appear with a uniform probability. However, the correct label is present with a probability of at least α.

The amount of noise in the aforementioned scenarios increases with the value of α and decreases with the value of β. In all the tested scenarios we constrained both parameters to sum to one.

Models In all cases, we compare the performance of our proposed methods with three baselines. First, we include the results of using the set of clean labels without added noise (`Superv`). This method gives us a lower bound on the error rate. Second, we show the expected error if the weak labels are used without any consideration (`Weak`). Third, we compare our approaches with optimistic superset loss (OSL) [9], as this is a popular technique to deal with weak labels (`OSL`).

As for our models, we explore three different scenarios. If we know the mixing matrix $\mathbf{M}$, we obtain the new virtual labels following the approach suggested in Section 3.3 (`Mproper`). If we do not know $\mathbf{M}$, and we assume a quasi Independent mixing matrix $\mathbf{M}$ we use the method suggested in Section 3.4 (`qIPL`). Finally, the alternative assumption of an Independent mixing matrix suggested in Section 3.4 is equivalent to the baseline `Weak`.

Implementation In order to evaluate the models we used the framework Keras that adds an abstraction layer on top of TensorFlow and Theano. Keras allows an easy specification of the loss function and automatic differentiation. In our case we used a Brier loss that is suitable for the virtual labels. All the models were trained with full batch gradient descent with a fixed learning rate of 1.0 for 40 epochs and with 10 times 10-fold cross-validation. For each dataset and mixing matrix $\mathbf{M}$ we trained a Logistic Regression (LR), and a Feed-fordward Neural Network (FNN) with two layers of 200 rectified linear units. Although all the comparisons on this section are focused on LR, the results of the FNN are shown in the second row of Figure 3 to illustrate the applicability of our method with richer hypothesis classes. All implementations are publicly available[3].

Results Fig. 3 shows the error rate of every model and mixing matrix $\mathbf{M}$ averaged over the 31 datasets, increasing the noise level from left to right; LR on top and FNN at the bottom. Although the mean error rate over several datasets is not fully informative, we performed a one tailed Wilcoxon rank sum test for each pair of models and every type of mixing matrix. The null hypothesis is that

[3] `https://github.com/Orieus/WeakLabelModel/`

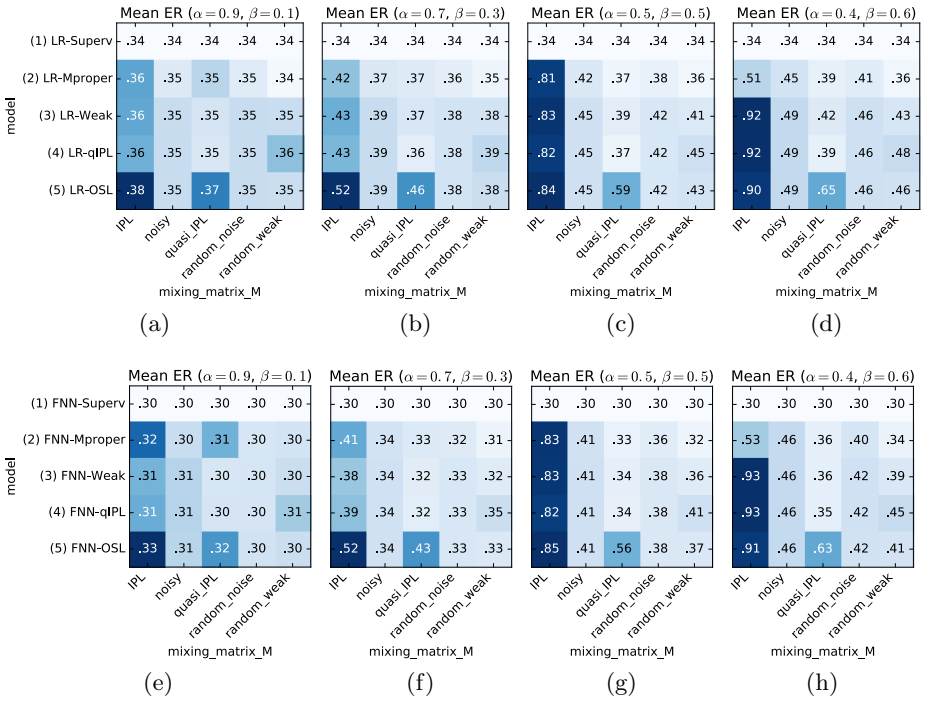

Fig. 3: Mean cross-validation error rates for 31 datasets.

the models follow the same distribution. All the conclusions are extracted from the statistical tests while the figures are only here to provide a visual intuition. Fig. 3a presents the case where the weak labels are similar to the true labels. Here the only significant comparison (p-value of 0.02) confirms that `Superv` outperforms `OSL`. In Fig. 3b for both the Independent Partial Labels (IPL) and quasi Independent Partial Labels (qIPL) mixing matrices, all methods surpass `OSL` (p-values of 10^{-5} and 10^{-4}, respectively). In Fig. 3c `Weak` and qIPL are still not statistically different but all the rest are (p-values under 10^{-3}). The IPL mixing matrix with this level of noise is a degenerate case, as the model that uses the known $\mathbf{M}$ can not retrieve the original labels. Fig. 3d shows how for higher noise levels, the differences between the models are highlighted and are easier to rank (p-values under 10^{-3} except for pairwise comparison between `Weak` and `qIPL` models which has a p-value of 0.31). Similar conclusions can be extracted from the second row of Fig. 3 for more complex FNNs.

As expected, if the random process that generated the noise is known, the best approach is to use the `Mproper` method. For that reason, when possible, it is advisable to try to estimate the mixing matrix from a clean set with true labels. When the mixing process is unknown, the results show that assuming a quasi independent or independent mixing process will help in most of the cases,

achieving the best results when the prior assumption is true. With respect to the poor performance of the `OSL` on the proposed scenarios, we hypothesise that this method makes important assumptions about particular correlations between classes with respect to the input space.

6 Conclusion

In this paper we built upon previous theoretical results to describe a simple approach to adapt existing classifiers that are based on the empirical minimization of proper or classification calibrated losses, in such a way that they explicitly incorporate the available information regarding the label quality. Furthermore, we show that the constructed weak loss achieves similar and in some cases surpasses the performance of a state-of-the-art method on a variety of scenarios.

References

1. Ambroise, C., Denoeux, T., Govaert, G., Smets, P.: Learning from an imprecise teacher: probabilistic and evidential approaches. In: Applied Stochastic Models and Data Analysis, vol. 1, pp. 100–105 (2001)
2. Angluin, D., Laird, P.: Learning from noisy examples. Machine Learning **2**(4), 343–370 (1988)
3. Cid-Sueiro, J.: Proper losses for learning from partial labels. In: Advances in Neural Information Processing Systems 25, pp. 1574–1582 (2012)
4. Cid-Sueiro, J., García-García, D., Santos-Rodríguez, R.: Consistency of losses for learning from weak labels. In: European Conference on Machine Learning and Knowledge Discovery in Databases, pp. 197–210. Springer Berlin Heidelberg (2014)
5. Cour, T., Sapp, B., Taskar, B.: Learning from partial labels. Journal of Machine Learning Research **12**, 1225–1261 (2011)
6. Grandvalet, Y.: Logistic regression for partial labels. In: 9th Information Processing and Management of Uncertainty in Knowledge-based System, pp. 1935–1941 (2002)
7. Grandvalet, Y., Bengio, Y.: Learning from partial labels with minimum entropy. Centre Universitaire de recherche en analyse des organizations (2004)
8. Hüllermeier, E., Beringer, J.: Learning from ambiguously labeled examples. Intell. Data Anal. **10**(5), 419–439 (2006)
9. Hüllermeier, E., Cheng, W.: Superset learning based on generalized loss minimization. In: European Conference on Machine Learning and Knowledge Discovery in Databases, pp. 260–275 (2015)
10. Jin, R., Ghahramani, Z.: Learning with multiple labels. In: Advances in Neural Information Processing Systems 15, pp. 897–904 (2002)
11. Nguyen, N., Caruana, R.: Classification with partial labels. In: SIGKDD international conference on Knowledge discovery and data mining, pp. 551–559 (2008)
12. Ni, Y., McVicar, M., Santos-Rodriguez, R., De Bie, T.: Understanding effects of subjectivity in measuring chord estimation accuracy. IEEE Transactions on Audio, Speech and Language Processing **21**(12), 2607–2615 (2013)
13. Raykar, V.C., Yu, S., Zhao, L.H., Valadez, G.H., Florin, C., Bogoni, L., Moy, L.: Learning from crowds. Journal of Machine Learning Research **99**, 1297–1322 (2010)
14. van Rooyen, B., Menon, A.K., Williamson, R.C.: Learning with symmetric label noise: The importance of being unhinged. In: Advances in Neural Information Processing Systems, pp. 10–18 (2015)