Adapting Linear Discriminant Analysis to the Paradigm of Learning from Label Proportions

M. Pérez-Ortiz
Dept. of Quantitative Methods
Universidad Loyola Andalucía
14004 Córdoba, Spain
Email: mariaperez@uloyola.es

P.A. Gutiérrez
Dept. of Computer Science
and Numerical Analysis
University of Córdoba
14071 Córdoba, Spain
Email: pagutierrez@uco.es

M. Carbonero-Ruz
Dept. of Quantitative Methods
Universidad Loyola Andalucía
14004 Córdoba, Spain
Email: mcarbonero@uloyola.es

C. Hervás-Martínez
Dept. of Computer Science
and Numerical Analysis
University of Córdoba
14071 Córdoba, Spain
Email: chervas@uco.es

***Abstract*—The recently coined term "learning from label proportions" refers to a new learning paradigm where training data is given by groups (also denoted as "bags"), and the only known information is the label proportion of each bag. The aim is then to construct a classification model to predict the class label of an individual instance, which differentiates this paradigm from the one of multi-instance learning. This learning setting presents very different applications in political science, marketing, healthcare and, in general, all fields in relation with anonymous data. In this paper, two new strategies are proposed to tackle this kind of problems. Both proposals are based on the optimisation of pattern class memberships using the data distribution in each bag and the known label proportions. To do so, linear discriminant analysis has been reformulated to work with non-crisp class memberships. The experimental part of this paper sets different objetives: 1) study the difference in performance, comparing our proposals and the fully supervised setting, 2) analyse the potential benefits of refining class memberships by the proposed approaches, and 3) test the influence of other factors in the performance, such as the number of classes or the bag size. The results of these experiments are promising, but further research should be encouraged for studying more complex data configurations.**

***Index Terms*—weak supervision, linear discriminant analysis, learning from label proportions**

I. Introduction

The new term weak supervision [1] originates from the need to tackle classification problems where the available information is, as the research area indicates, weak, and not as accessible as in a standard supervised classification problem (where a label is associated to each pattern). An example is the problem of learning a model using class-probabilities [2]. This is, for each pattern the known information is a vector indicating the pattern probability of belonging to each of the classes of the problem, and the final aim is to construct a prediction model from these probabilities. Another example with increased popularity is the semi-supervised learning setting [3], where both labelled and unlabelled information is used to construct the predictive model.

In this sense, the recently term known as learning from label proportions (which can be framed under the notion of weak supervision), has emerged and is receiving attention from the machine learning research community [4], [5], [6]. This paradigm study those learning problems where training data are given as disjoint groups of patterns, and the only known information are the proportions of the labels for each group or bag. As in supervised classification, the objective remains that of predicting the label for unseen individual instances, in contrast to multi-instance learning, whose objective is to predict a single label for a whole new bag. A wide range of different applications can be found in the literature concerning this type of problems, specially concerning anonymous data, non-monitoring processes and all data with a black-box nature. A clear example would be the creation of a system for poll prediction (i.e. predicting what a citizen is going to vote). Under this setting, the most commonly available data are those obtained from polling stations, i.e. groups of citizens, where we only have the proportion of their voting, rather than the preferred party of each of these citizens.Although semi-supervised problems have been extensively studied, the paradigm from learning from label proportions (or aggregate outputs) remains mostly unexplored [4], [5], [6], [7], [8], [9], [10]. More specifically, probabilistic classifiers have been considered [4], [5], [7], given that they work naturally with pattern class-probabilities. Other algorithms have been adapted as well, such as neural networks [8] or support vector machines [8], [10]. One of these works have derived a complexity upper bound for binary problems [5], proving that the complexity scales exponentially with respect to the bag sample size. Furthermore, it has been seen that the performance of such algorithms decreases with the increase of the number of classes in the problem. Despite the promising results of these works, further research should be performed in this line, to adapt the rest of classifiers in the literature and explore other hypotheses.

In this paper, a set of simple approaches are proposed based on discriminant analysis [11] (being therefore useful for classification and supervised dimensionality reduction), although different classifiers could also be used. The main objective of our proposals is to refine (iteratively) a priori class-probabilities (estimated through the given label proportions) based on two sources: 1) pattern distribution or, alternatively, 2) estimated classifier probabilities. Instead of tackling this problem using a transductive approach (i.e. trying to estimate the original class labels before the learning process), we reformulate a well-known classification strategy (discriminant

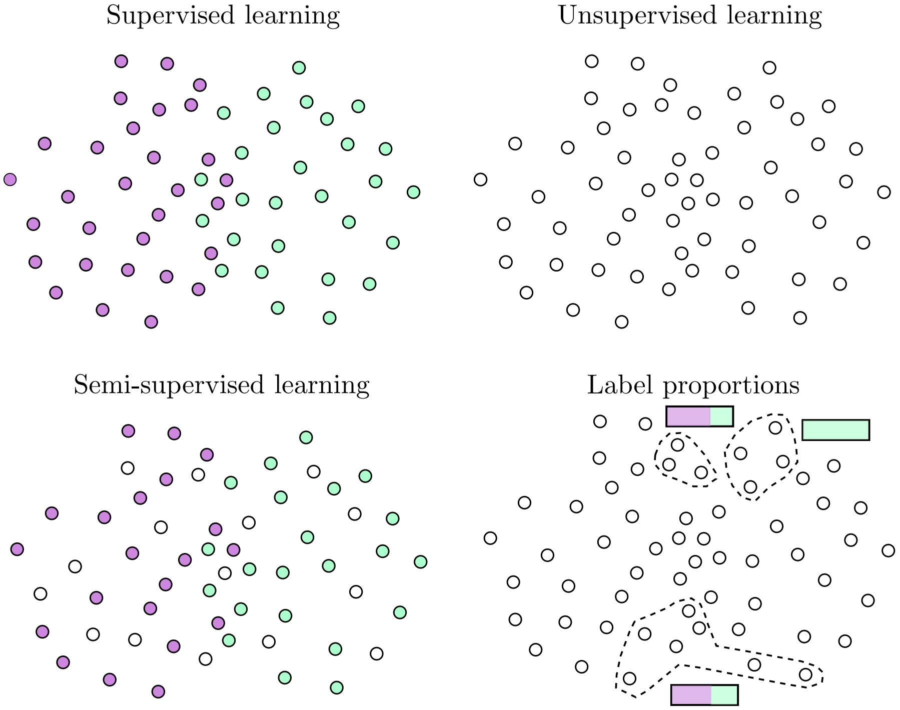

Fig. 1: Representation of the available information for four different learning paradigms. Each colour indicates a class (and non-coloured patterns represent unsupervised patterns). In the case of label proportions the data are grouped into different bags of patterns and the only supervised available information are the label proportions of each bag (represented by a coloured rectangle). Note that, for visual simplicity, only some patterns have been grouped in a bag in the case of label proportions.

analysis) to deal with class-uncertainty. As stated before, two approaches are derived: 1) a filter-based approach, based on distances, and 2) a wrapper one, which uses the classifier itself to approximate probabilities. A thorough set of experiments is performed, using 28 benchmark datasets: 14 binary and 14 multiclass, with up to 21 classes (note that the multiclass setting has been barely studied in this literature [6], [5], [4], [10]). Our experiments try to answer experimentally to different research questions: Firstly, is it possible to construct a valid prediction tool, when we are only given bags of data and the associated label proportions of each group (even when no individual datum is labelled)?. Secondly, could this approach compare to the fully supervised case, obtaining similar results?. And finally, does the data complexity (in this case measured by the bag size and number of classes) poses a serious handicap in this learning area?.

The rest of this paper is organised as follows: Section 2 includes previous notions and the description of the proposed methodologies; Section 3 presents the experimental part and analyses the results; and finally, Section 4 outlines the conclusions and future work.

II. Methodology

In the standard supervised classification setting the objective is to assign an input vector $\mathbf{x} \in \mathbb{R}^m$ to one of K classes c_k, where $k \in \{1, \ldots, K\}$, where N be the number of sample patterns and N_k is the number of samples of the k-th class. The final aim is to compute a function $f : X \to Y$ using a sample $D = \{\mathbf{x}^i, y^i\}_{i=1}^N \in X \times Y$.

Concerning the learning from label proportions setting, suppose, as said, a dataset (composed of N data) and divided into b bags $D = B_1 \cup B_2 \cup \ldots \cup B_b$, where $B_i \cap B_j = \emptyset$, $\forall i \neq j$. A bag $B_i = \{\mathbf{x}^{i1}, \mathbf{x}^{i2}, \ldots, \mathbf{x}^{iN_i}\}$ groups N_i patterns and contains the only supervised information: the counts C_{ik} which correspond to the number of patterns in B_i that belong to class c_k. Similarly, bag class information can be given in terms of proportions, $p_{ik} = \frac{C_{ik}}{N_i} \in [0, 1]$ with $\sum_{c_k \in Y} p_{ik} = 1$.

Fig. 1 shows a graphical comparison between this learning paradigm and other common ones. For visual simplicity, in the case of label proportions not all the patterns are grouped. As stated, the only supervised information known is the label proportions per bag (shown in the figure by different coloured proportions in the rectangle associated to each bag). The difficulty in this case resides in assigning each instance to its class, which can be thought as uncertainty of the label proportions. Since each group has a given label proportion and could involve a different amount of uncertainty, two type of bags could be defined: full bags (with zero uncertainty, composed only of patterns from the same label) and non-full bags (with a mix of patterns from different classes). These terms will be used throughout the paper. Note that in the learning from label proportions case represented in Fig. 1 there are two non-full bags and one full bag.

The technique proposed in this paper considers the use of Linear Discriminant Analysis (LDA) [11].In this case, although originally LDA does not consider probabilities, its formulation can be easily changed to take them into account. The rest of this section describes the fully-supervised LDA and presents the proposed method.

A. Fully-supervised LDA

This is one of the pioneer and leading techniques in machine learning, being both used in dimensionality reduction and classification [11]. The objective is to compute the optimum linear projection for the data (the one which separates the classes in the best way possible). To do this, this technique considers two objectives: the maximisation of the between-class distance and the minimisation of the within-class distance, by the use of covariance matrices ($\mathbf{S}_{\mathrm{b}}$ and $\mathbf{S}_{\mathrm{w}}$, respectively) and the so-called Rayleigh coefficient ($J(\boldsymbol{\beta}) = \frac{\boldsymbol{\beta}^T \mathbf{S}_b \boldsymbol{\beta}}{\boldsymbol{\beta}^T \mathbf{S}_w \boldsymbol{\beta}}$, where $\boldsymbol{\beta}$ is the projection). To achieve this, the K leading eigenvectors, associated to the highest eigenvalues of $\mathbf{S}_{\mathrm{w}}^{-1} \cdot \mathbf{S}_{\mathrm{b}}$, are calculated, and these compose the projection function, which can be later used as a discriminant or visualisation technique.

B. Weakly-supervised LDA

As said, LDA represents the class distributions using covariance matrices. In order to combine label proportions with this classifier and assist this representation, the class means and within-class covariance matrix can be rewritten as follows:

$$\mu_k = \frac{\sum_{i=1}^{b} \sum_{j=1}^{N_i} p_{ik} \mathbf{x}^{ij}}{\sum_{i=1}^{b} p_{ik} \cdot N_i}, \ k = 1, \ldots, K, \tag{1}$$

$$\mathbf{S}_{\mathrm{w}} = \frac{1}{p} \sum_{k=1}^{K} \sum_{i=1}^{b} \sum_{j=1}^{N_i} p_{ik} [\mathbf{x}^{ijk} - \mu_k][\mathbf{x}^{ijk} - \mu_k]^T, \tag{2}$$

where $\mathbf{p}_{ij} = \{p_{i1}, \ldots, p_{iK}\}$, $p = \sum_{k=1}^{K} \sum_{i=1}^{b} p_{ik} \cdot N_i$, and p_{ik} is the a priori class probability for B_i and c_k (obtained from the initial label proportions). By this approach, we directly consider that all the data in this bag have the same probability (p_{ik}). This formulation will allow patterns with a higher membership for class c_k to contribute to a larger extent to the computation of the mean and within-class covariance of the class. The technique here described is referred in the experiments to as Simple Weighting Scheme (SWS-LDA).

C. A distance-based filter approach to weakly-supervised LDA

Label proportions are generally used as a first estimation of the class probabilities per instance. In this sense, our proposal tries to improve this estimation and not assume the same that probability is equal for all the bag data. Our first idea is that these class probabilities can be iteratively improved using on data distribution in the input space. To this end, the method here proposed relies on the previously presented definition of class-mean (Eq. 1) and makes use of distance relations. One of the crucial characteristics of this proposal is that the original label ratios are maintained (although individual proportions change), since this is the only true supervised information available.

For a graphical description of the main idea behind this technique see Fig. 2, where a hypothetical bag B_1 of three patterns is represented. It can be seen that the pattern that will have the greatest uncertainty will be $\mathbf{x}^{12}$, since it lies on the intersection of c_1 and c_2. For this example, the original label proportions are $p_{11} = \frac{2}{3}$ and $p_{12} = \frac{1}{3}$ (i.e. two patterns within B_1 belong to c_1 and one of them to c_2). It is clear that these initial probabilities can be improved based on the data distribution (specially for $\mathbf{x}^{11}$ and $\mathbf{x}^{13}$). In the absence of prior information, x^{12} (which lies in an overlapping region between c_1 and c_2), would have the same probability of belonging to both classes. However, as the original label proportions are considered in the refinement step, $\mathbf{x}^{12}$ would be assigned in this case to c_1.

The first step of our approach is the computation of the means per class using Eq. 1, where each pattern $\mathbf{x}^{ij}$ associated to bag B_i contributes with a weight of p_{ik} to the computation of μ_k. In this sense, μ_k estimates the real value of the mean for class c_k. The second step is to compute the relative distance of each pattern to the mean of each class using the previous approximated means:

$$P(y^{ij} = c_k | \mathbf{x}^{ij}) = p^*_{ijk} = 1 - \frac{d(\mathbf{x}^{ij}, \mu_k)}{\sum_{z=1}^{K} d(\mathbf{x}^{ij}, \mu_z)}, \tag{3}$$

where $j = \{1, \ldots, N_i\}$, $k = \{1, \ldots, K\}$, and d represents a distance relation (in our case the Euclidean distance). The vector $\mathbf{p}^*_{ij} = \{p_{ij1}, \ldots, p_{ijK}\}$ provides information about the relative closeness of $\mathbf{x}^{ij}$ to all the classes, where $\sum_{k=1}^{K} p^*_{ijk} = 1$.

The next step is to refine the associated probabilities based on this distance relation. Note that it is crucial to maintain the original ratios p_{ik} (as this is the only true information about the labelling). To maintain these ratios while optimising the different weights per pattern, the following formulation can be used:

$$w_{ijk} = \frac{p^*_{ijk}}{\sum_{z=1}^{N_i} p^*_{izk}} \cdot C_{ik}, \tag{4}$$

where the first term represents a ranking of relative distances of the patterns inside bag B_i, and w_{ijk} represents the associated weight of pattern $\mathbf{x}^{ij}$ with respect to class c_k. Note this formulation ensures $\frac{p^*_{ijk}}{\sum_{z=1}^{N_i} p^*_{izk}} = 1, \ k = 1, \ldots, K$, and thus, when multiplying this by C_{ik}, the original label proportions are maintained.

For the sake of understanding, analyse the computation of these weights for the example in Fig. 2:

$$\begin{array}{lll} p^*_{111} = 0.750, & p^*_{121} = 0.516, & p^*_{131} = 0.302 \\ p^*_{112} = 0.250, & p^*_{122} = 0.484, & p^*_{132} = 0.698 \\ w_{111} = 0.957, & w_{121} = 0.658, & w_{131} = 0.385 \\ w_{112} = 0.175, & w_{122} = 0.338, & w_{132} = 0.487 \end{array}$$

In this case, it can be seen that the initial class probabilities (derived from the label proportions) are refined. $\mathbf{x}^{11}$ has increased its weight for c_1 and $\mathbf{x}_{13}$ for c_2 (note that initial class probabilities for this pattern were $p_{11} = \frac{2}{3}$ and $p_{12} = \frac{1}{3}$). In

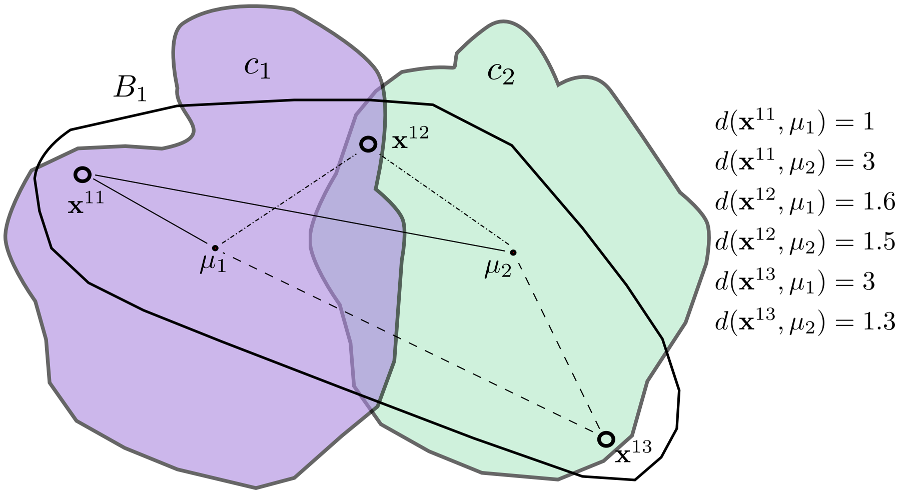
B_1
c_1
c_2
$\mathbf{x}^{11}$
$\mathbf{x}^{12}$
$\mathbf{x}^{13}$
μ_1
μ_2
$d(\mathbf{x}^{11}, \mu_1) = 1$
$d(\mathbf{x}^{11}, \mu_2) = 3$
$d(\mathbf{x}^{12}, \mu_1) = 1.6$
$d(\mathbf{x}^{12}, \mu_2) = 1.5$
$d(\mathbf{x}^{13}, \mu_1) = 3$
$d(\mathbf{x}^{13}, \mu_2) = 1.3$

Fig. 2: Representation of the main idea behind the proposal for a toy example where $p_{11} = \frac{2}{3}$ and $p_{12} = \frac{1}{3}$. In this case, these initial class probabilities can be further improved computing distances to class centroids, which results in $\mathbf{x}^{11}$ and $\mathbf{x}^{12}$ having a greater probability of belonging to c_1 and $\mathbf{x}^{13}$ to c_2.

the case of $\mathbf{x}^{12}$, although the original distance relation should indicate that this pattern is closer to c_2, our methodology is able to fix this.

By this strategy, the initial probabilities p_{ik} (for pattern $\mathbf{x}^{ij}$) are updated based on the classes structure, and a new estimated weight w_{ijk} is obtained. This process is made iterative, refining the class means using these weights until convergence as follows:

$$\hat{\mu}_k = \frac{\sum_{i=1}^{b} \sum_{j=1}^{N_i} w_{ijk} \mathbf{x}^{ij}}{\sum_{i=1}^{b} w_{ik} \cdot N_i}. \tag{5}$$

where $w = \sum_{i=1}^{b} \sum_{j=1}^{N_i} \sum_{k=1}^{K} w_{ijk}$. Fig. 3 shows the pseudocode of this distance-based weight refinement procedure, where $\mathbf{C} = \{C_{ik}; i = 1, \ldots, b; k = 1, \ldots, K\}$.

Pseudocode for Filter Weighting Scheme (FWS)

```
• Input: training data $D$, label proportions $\mathbf{C}$.
• Output: pattern-weight per class $\mathbf{w}$.
  $t \leftarrow 1$
  while not stopping criterion
      1. if $t = 1$
            Compute $\boldsymbol{\mu}$ using $D$ and $\mathbf{p}$, Eq. (1).
         else
            Compute $\boldsymbol{\mu}$ using $D$ and $\mathbf{w}$, Eq. (5).
         end
      2. Compute pattern distances to $\boldsymbol{\mu}$: $d(\mathbf{x}^{ij}, \boldsymbol{\mu})$.
      3. Rank distances using Eq. 3 ($\mathbf{p}^*$).
      4. Compute $\mathbf{w}$ using Eq. 4, $\mathbf{p}^*$ and $\mathbf{C}$.
      6. $t \leftarrow t + 1$
  end while
```

Fig. 3: Different steps considered for the Filter Weighting Scheme to obtain the final weights per pattern.

After applying this procedure, we obtain a set of optimised weights. At this point different classifiers could be used. However, we use the well-known LDA algorithm, because of its natural extension to this paradigm. The class mean and covariance matrices used in LDA are computed in this case using the estimated weights using Eq. (5) and:

$$\hat{\mathbf{S}}_{\mathrm{w}} = \frac{1}{w} \sum_{k=1}^{K} \sum_{i=1}^{b} \sum_{j=1}^{N_i} w_{ijk} [\mathbf{x}^{ijk} - \mu_k][\mathbf{x}^{ijk} - \mu_k]^T. \tag{6}$$

The projection $\boldsymbol{\beta}$ is computed solving an eigenvector problem via the Rayleigh quotient. The technique here presented is referred in the experiments to as Filter Weighting Scheme (FWS-LDA).

D. A wrapper approach to weakly-supervised LDA

Class-memberships could also be refined using weak classifiers, which learn a model based on weak information. More specifically, the classifier can be constructed in each iteration using the estimated weights $\mathbf{w}$. Class-probabilities could be then estimated from the obtained model (for LDA using the projection and a soft-max function). These probabilities are used to further refine $\mathbf{w}$ considering the number of instances in B_i that belong to class label c_k (i.e., C_{ik}).

In the same vein than in Sections II-B and II-C, the strategy here starts computing the class means using Eq. (1), where each pattern $\mathbf{x}^{ij}$ of group B_i contributes with a weight of p_{ik} to the computation of μ_k. Apart from these means per class, the corresponding LDA model is trained as before, using the associated covariance matrix computed by Eq. (2) and the Rayleigh coefficient.

Once that an initial $\boldsymbol{\beta}$ has been obtained, the probability that pattern $\mathbf{x}^{ij}$ has of belonging to class c_k can be estimated from the projection using the soft-max function:

$$P(y^{ij} = c_k | \mathbf{x}^{ij}) = p^*_{ijk} = \frac{e^{((\mathbf{x}^{ij})^{\mathrm{T}} \boldsymbol{\beta}_k)}}{\sum_{z=1}^{K} e^{((\mathbf{x}^{ij})^{\mathrm{T}} \boldsymbol{\beta}_z)}}. \tag{7}$$

In the next phase, these probabilities will be refined using the initial proportions applying Eq. (4). This refinement is captured in the associated weights $\mathbf{w}$, which are then used again for constructing the mean per class, the covariance matrix and a new LDA model, using Eqs. (5) and (6). This process of computing the LDA projection using the pattern weights and estimating the new weights using the soft-max function and label proportions per bag is also repeated until convergence. Fig. 4 shows the pseudocode of this classifier-based weight refinement procedure, where $\mathbf{C} = \{C_{ik}; i = 1, \ldots, b; k = 1, \ldots, K\}$.

Pseudocode for the Wrapper Weighting Scheme (WWS-LDA)

- **Input**: training data D, label proportions $\mathbf{C}$.
- **Output**: pattern-weight per class $\mathbf{w}$.

```
$t \leftarrow 1$
while not stopping criterion
    1. if $t = 1$
           Compute $\boldsymbol{\mu}$ using $D$ and $\mathbf{p}$, Eq. (1).
       else
           Compute $\boldsymbol{\mu}$ using $D$ and $\mathbf{w}$, Eq. (5).
       end
    2. Compute LDA projection $\boldsymbol{\beta}$.
    3. Estimate probabilities using Eq. 7 ($\mathbf{p}^*$).
    4. Compute $\mathbf{w}$ using Eq. 4, $\mathbf{p}^*$ and $\mathbf{C}$.
    5. $t \leftarrow t + 1$
end while
```

Fig. 4: Different steps considered for the Wrapper Weighting Scheme to obtain the final weights per pattern.

The method here proposed is referred in the experiments to as Wrapper Weighting Scheme (WWS-LDA).

III. Experiments

The experimental design and the results obtained are presented in this section.

The most common relevant source of information for learning from label proportion is private data. Consequently, there are not publicly available datasets. In this way, synthetic or classical supervised datasets are usually considered in the literature, after transforming them into label proportions with the objective of validating the algorithms.

A set including 28 supervised benchmark datasets (with 14 binary classification problems and 14 multi-class ones, some of them with more of 20 classes) are used to compare the two approaches proposed in this paper. Different bag sizes are tested (more specifically, 3, 5 and 10 patterns per bag) to stablish conclusions in the influence of this parameter in the classifier performance. The bags are randomly constructed, all algorithms being applied to same bags. The standard supervised algorithm is also compared to learning from label proportions, in order to check the consequences of weekly supervised information in the final performance of the classifiers.

TABLE I: Characteristics of the datasets used for the experiments: number of patterns (N), features (m), classes (K) and baseline mean and standard deviation Acc LDA fully supervised results.

Dataset	N	m	K	LDA Acc
Binary datasets				
appendicitis (AP)	106	7	2	86.73 ± 9.40
bands	365	19	2	66.31 ± 4.36
breast	286	15	2	70.28 ± 5.96
card	690	51	2	86.38 ± 4.17
colic	368	60	2	82.23 ± 3.88
credit-a	690	43	2	85.77 ± 4.72
haberman	306	3	2	74.19 ± 4.94
heart	270	13	2	82.96 ± 5.00
hepatitis	155	19	2	84.54 ± 6.28
housevotes	232	16	2	96.96 ± 3.58
mammographic	830	5	2	80.36 ± 4.65
pima	768	8	2	77.60 ± 4.29
sick	3772	33	2	95.57 ± 0.90
wisconsin	683	9	2	96.04 ± 1.84
Multiclass datasets				
cleveland	297	13	5	58.97 ± 7.60
dermatology	358	34	6	96.07 ± 3.58
ecoli	336	7	8	78.29 ± 4.62
flare	1066	38	6	75.70 ± 3.24
glass	214	9	6	45.78 ± 7.66
hayes-roth	160	4	3	54.38 ± 14.15
horse	364	58	3	65.16 ± 8.38
hypothyroid	3772	33	4	93.03 ± 0.28
iris	150	4	3	98.00 ± 4.50
nursery	12960	26	5	91.13 ± 0.80
primary-tumor	339	23	21	25.68 ± 5.09
soybean	683	84	19	94.14 ± 2.94
wine	178	13	3	98.86 ± 2.41
zoo	101	21	7	95.00 ± 5.27

Different methods haven been compared in this paper:

- A standard fully supervised version of Linear Discriminant Analysis (LDA), where the original labels are used for learning. Although this information is not available for learning from label proportions, these results are considered as a baseline to study whether the paradigm of the algorithms developed might yield similar results to the complete supervised setting in this case.
- Simple Weighting Scheme combined with the LDA method (SWS-LDA). This method is described in Section II-B, the class weights for each instance being fixed using the initial label proportions. Depending on the bag size, the results of this method are given three different acronyms: SWS-LDA-3, SWS-LDA-5 and SWS-LDA-10.
- The method Filter Weighting Scheme combined with the LDA technique (FWS-LDA) is described in Section II-C, the class weights being optimised by a distance-based strategy. The results for this method are referred to as FWS-LDA-3, FWS-LDA-5 and FWS-LDA-10, depending on the bag size.
- Finally, a Wrapper Weighting Scheme combined with the

LDA technique (WWS-LDA) is considered. The methodology is the one introduced in Section II-D, the class weights being optimised by a classifier-based strategy. In this case, the results are referred to as WWS-LDA-3, WWS-LDA-5 and WWS-LDA-10.

The only parameter for FWS and WWS methods is the stop criterion. The algorithms is considered to have converged when the total change for the estimated weights is less than 10^{-5}.

In order to measure the performance of the different classifiers obtained, we consider the accuracy (Acc), i.e. the percentage of correctly classified patterns:

$$Acc = 100 \cdot \frac{1}{N} \sum_{i=1}^{N} [\![\hat{y}_i = y_i]\!], \tag{8}$$

where $[\![\cdot]\!]$ is the indicator function (being 1 if the condition is true, and 0 otherwise) and $\hat{y}_i$ is the predicted target for $\mathbf{x}_i$.

A 10-fold partition procedure is used for the experimental design. TABLE I presents the characteristics of the 28 datasets, including the number of instances (N), input variables (m) and classes (K). The fully supervised LDA results can also be found in this table. Note that we are considering multiclass datasets with up to 21 classes, which poses a serious handicap for the setting of learning from label proportions.

TABLE II shows the complete set of results, the best performing algorithm for each bag size being marked in bold face, while the second one is marked in italics. Now, different issues are discussed from this table.

If we start by comparing the values obtained for FWS-LDA-3 to the results of fully supervised LDA in TABLE I, it is clear that, generally, the fully supervised setting leads to results which are better or similar to those of the label proportions proposal. In binary datasets, similar performance is obtained for six datasets (considering a range of 1% accuracy), the performance is better in seven cases (the difference being larger than 1% of accuracy), and it is worse for one dataset (the range being larger than 1%). When considering the same ranges in the multiclass datasets, supervised LDA obtains similar results in four datasets, better in nine datasets and worse in one.

If we compare WWS-LDA-3 to the supervised LDA method, when analysing binary datasets, LDA and WWS-LDA-3 are similar in five cases (range within 1%), LDA is better in seven datasets (range larger than 1%), and it is worse for two datasets. The same comparison for multiclass datasets, using the same ranges, establishes that the supervised approach is similar in seven datasets, better in six cases and worse in one.

Taking into account the difficulty inherent to learning from label proportions, the results are acceptable when compared to the supervised version of the algorithms proposed. However, there is still room for improvement, what makes clear that the proposals of this paper and other existing approaches need to be improved (given the additional problems involved in multi-class datasets). In any case, the methodologies presented in this paper seem to reconstruct the information about the original labels for many of the datasets by using only weakly-supervised information.

On the other hand, by examining TABLE II, we can conclude that the process of adjusting the class probabilities by the iterative methods (FWS-LDA and WWS-LDA) improves the results with respect to considering only the initial probabilities (SWS-LDA) independently of the bag sizes (FWS-LDA being better than SWS-LDA for 58 datasets out of 84, and WWS-LDA wining in 53 datasets). As an example, consider the significant difference of performance for iris and wine datasets. Moreover, WWS-LDA seems to obtain slightly better results than FWS-LDA, WWS-LDA obtaining the best performance of all methods for 50 out of 84 datasets. Finally, when more difficult problems are considered (i.e. when the size of the bags is larger or when the number of classes is increases), FWS-LDA seem to yield similar results to SWS-LDA, the results LDA and WWS-LDA being significantly better. This is very clear for some datasets such as zoo. The class probabilities are harder to be optimised with this kind of datasets for all methods which learn a classifier from only label proportions, specially if only a filter approach is considered. However, the wrapper method (WWS-LDA) leads to good performance for large bag sizes and for datasets with many classes: the difference of performance favouring WWS-LDA with respect to FSW-LDA and SSW-LDA is very high for wine, soy-bean, and iris datasets, when using 5 and 10 patterns per bag.

To conclude the analysis of results, TABLE II also includes the mean results and rankings for all methods and configurations considered (different bag sizes and number of classes). From these results, it is clear that the filter and wrapper methods perform quite similarly for binary classification problems. Nonetheless, WWS-LDA shows better results when considering more complex configurations (larger bag sizes and multiple classes), where the difference of performance is $12,25\%$ on average for multiclass datasets and bags of 10 instances. Note also that WWS-LDA involves a higher computational cost.

IV. Conclusions

The approaches developed in this paper are contextualised on the topic of learning from label proportions, where the supervised information available is the label proportion for each bag of data. The methods here proposed make use of linear discriminant analysis and are based on an iterative refinement of class probabilities, one of them based on the position of patterns in the input space (acting as a filter approach for LDA) and the other based on the estimated probabilities of LDA (wrapper approach). The final model is constructed using these refined weights to estimate the class distributions (i.e. mean per class and covariance matrices). Our experiments have shown that our approaches improve the result, leading to a promising accuracy even when considering multiclass datasets. Moreover, our experiments have also proven that the complexity of such problems grows with the bag size and number of classes.

TABLE II: *Acc* test experimental results obtained (in mean and standard deviation).

Binary	SWS-LDA-3	FWS-LDA-3	WWS-LDA-3	SWS-LDA-5	FWS-LDA-5	WWS-LDA-5	SWS-LDA-10	FWS-LDA-10	WWS-LDA-10
appendicitis	79.27 ± 3.55	*85.82 ± 8.36*	**89.55 ± 8.62**	80.18 ± 2.77	**83.91 ± 4.70**	*83.09 ± 10.46*	80.18 ± 2.77	*83.00 ± 5.88*	**83.91 ± 9.21**
bands	**65.47 ± 2.70**	**65.47 ± 2.99**	*62.74 ± 7.94*	**63.83 ± 1.27**	**63.83 ± 1.80**	*57.24 ± 6.37*	**63.30 ± 1.02**	**63.30 ± 1.02**	*58.09 ± 7.29*
breast	*70.30 ± 1.43*	**71.70 ± 2.82**	68.17 ± 6.79	**70.28 ± 1.69**	**70.28 ± 1.69**	*68.84 ± 7.30*	**70.64 ± 1.49**	*70.30 ± 1.43*	67.44 ± 8.67
card	*84.49 ± 6.07*	84.20 ± 5.89	**85.65 ± 3.58**	75.51 ± 4.71	*76.52 ± 5.15*	**84.93 ± 3.50**	61.01 ± 3.51	*62.61 ± 4.03*	**85.36 ± 3.45**
colic	72.02 ± 6.20	*73.91 ± 5.44*	**81.25 ± 4.52**	66.88 ± 6.94	*67.95 ± 6.13*	**75.03 ± 8.92**	64.41 ± 2.89	*65.50 ± 3.46*	**73.37 ± 11.80**
credit-a	81.88 ± 3.07	*82.32 ± 3.33*	**84.78 ± 4.39**	76.67 ± 5.53	*79.13 ± 4.98*	**84.64 ± 4.69**	62.75 ± 4.21	*64.93 ± 3.73*	**84.93 ± 5.03**
haberman	73.20 ± 1.25	*73.54 ± 2.25*	**76.15 ± 4.03**	73.53 ± 1.00	*73.86 ± 1.40*	**74.19 ± 4.91**	73.53 ± 1.00	*73.86 ± 1.40*	**73.85 ± 4.38**
heart	*77.78 ± 7.20*	**82.59 ± 4.95**	77.41 ± 8.63	69.63 ± 7.57	**76.67 ± 6.06**	*76.30 ± 8.41*	61.85 ± 3.51	*70.37 ± 8.00*	**73.33 ± 10.00**
hepatitis	*80.63 ± 4.36*	**81.29 ± 4.74**	74.79 ± 7.69	*80.00 ± 1.86*	**81.29 ± 4.74**	75.38 ± 7.68	*79.38 ± 2.38*	*79.38 ± 2.38*	**80.13 ± 11.35**
housevotes	93.48 ± 6.87	*95.22 ± 4.78*	**96.96 ± 3.58**	90.07 ± 6.81	*91.38 ± 4.59*	**96.96 ± 3.58**	75.85 ± 9.34	*89.22 ± 6.87*	**95.65 ± 5.80**
mammographic	**80.72 ± 4.29**	*80.60 ± 4.15*	78.67 ± 4.44	**80.72 ± 4.95**	*80.60 ± 4.11*	77.71 ± 2.91	*78.43 ± 6.26*	**80.24 ± 4.41**	77.11 ± 3.01
pima	70.06 ± 2.88	**74.22 ± 4.17**	*73.31 ± 3.06*	65.50 ± 1.17	*68.62 ± 1.68*	**70.19 ± 5.83**	64.98 ± 0.59	*65.37 ± 0.84*	**67.32 ± 4.10**
sick	**93.88 ± 0.08**	**93.88 ± 0.08**	*92.58 ± 2.06*	**93.88 ± 0.08**	**93.88 ± 0.08**	*90.27 ± 2.51*	**93.88 ± 0.08**	**93.88 ± 0.08**	*88.23 ± 2.70*
wisconsin	85.79 ± 3.96	*95.31 ± 2.06*	**95.46 ± 1.61**	75.10 ± 4.37	*95.16 ± 1.56*	**95.32 ± 1.80**	65.01 ± 0.48	**95.31 ± 1.67**	*95.02 ± 1.42*
Mean	79.21	**81.43**	*81.25*	75.84	*78.79*	**79.29**	71.09	*75.52*	**78.84**
Ranking	2.36	**1.71**	*1.93*	2.46	**1.68**	*1.86*	2.54	*1.75*	**1.71**
Multiclass	SWS-LDA-3	FWS-LDA-3	WWS-LDA-3	SWS-LDA-5	FWS-LDA-5	WWS-LDA-5	SWS-LDA-10	FWS-LDA-10	WWS-LDA-10
cleveland	54.22 ± 1.22	*54.56 ± 1.85*	**57.57 ± 6.11**	**53.89 ± 0.89**	**53.89 ± 0.89**	*50.86 ± 9.18*	**53.89 ± 0.89**	**53.89 ± 0.89**	*49.79 ± 7.54*
dermatology	91.61 ± 3.76	*93.29 ± 3.78*	**96.07 ± 3.58**	85.74 ± 7.52	*87.12 ± 8.10*	**96.36 ± 2.98**	53.09 ± 4.52	*56.15 ± 4.69*	**95.23 ± 4.04**
ecoli	56.85 ± 4.95	*58.32 ± 4.67*	**75.00 ± 4.00**	48.52 ± 5.11	*51.24 ± 5.83*	**71.69 ± 4.49**	*43.15 ± 2.24*	*43.15 ± 2.24*	**62.21 ± 5.33**
flare	75.33 ± 3.33	*75.51 ± 3.29*	**75.61 ± 3.28**	71.67 ± 4.20	*72.51 ± 3.74*	**75.23 ± 3.47**	37.23 ± 2.98	*38.26 ± 3.74*	**74.30 ± 3.29**
glass	**32.73 ± 0.78**	**32.73 ± 0.78**	**32.73 ± 0.78**	**32.73 ± 0.78**	**32.73 ± 0.78**	**32.73 ± 0.78**	**32.73 ± 0.78**	**32.73 ± 0.78**	**32.73 ± 0.78**
hayes-roth	45.63 ± 12.52	*47.50 ± 12.22*	**60.00 ± 9.41**	44.38 ± 9.06	*45.00 ± 10.54*	**51.25 ± 9.22**	35.63 ± 8.36	*38.13 ± 10.40*	**53.75 ± 12.22**
horse	**67.33 ± 6.31**	**67.33 ± 6.13**	*60.15 ± 9.41*	**63.48 ± 4.74**	*62.39 ± 5.15*	54.69 ± 7.96	**61.82 ± 1.69**	**61.82 ± 1.69**	*58.88 ± 10.55*
hypothyroid	92.52 ± 0.63	**92.55 ± 0.83**	*88.34 ± 2.30*	**92.50 ± 0.56**	*92.42 ± 0.59*	89.98 ± 3.37	**92.34 ± 0.34**	**92.34 ± 0.34**	*88.34 ± 2.62*
iris	84.00 ± 10.04	*85.33 ± 11.24*	**98.00 ± 4.50**	82.67 ± 11.42	*85.33 ± 10.80*	**98.00 ± 4.50**	77.33 ± 7.17	*84.00 ± 9.53*	**98.00 ± 4.50**
nursery	*90.24 ± 0.71*	**90.25 ± 0.75**	89.97 ± 0.84	89.43 ± 0.72	*89.48 ± 0.70*	**88.97 ± 1.13**	86.39 ± 0.65	*86.50 ± 0.60*	**87.99 ± 0.98**
primary-tumor	**24.79 ± 1.64**	**24.79 ± 1.64**	**24.79 ± 1.64**	**24.79 ± 1.64**	**24.79 ± 1.64**	**24.79 ± 1.64**	**24.79 ± 1.64**	**24.79 ± 1.64**	**24.79 ± 1.64**
soybean	91.94 ± 2.45	*92.08 ± 2.72*	**93.26 ± 2.54**	89.15 ± 3.30	*89.30 ± 3.36*	**93.41 ± 2.80**	*80.23 ± 3.23*	80.08 ± 4.26	**90.62 ± 2.24**
wine	89.90 ± 6.82	*92.71 ± 5.33*	**99.41 ± 1.86**	86.54 ± 11.30	*93.89 ± 6.65*	**98.86 ± 2.41**	78.69 ± 9.32	*87.68 ± 7.73*	**98.86 ± 2.41**
zoo	79.09 ± 12.07	*80.09 ± 12.56*	**94.00 ± 5.16**	59.36 ± 8.88	*62.36 ± 12.28*	**89.09 ± 11.98**	46.55 ± 6.68	*47.55 ± 6.33*	**83.09 ± 15.72**
Mean	69.73	*70.50*	**74.64**	66.06	*67.32*	**72.57**	57.42	*59.08*	**71.33**
Ranking	2.61	*1.82*	**1.57**	2.39	*1.89*	**1.71**	2.43	*2.00*	**1.57**

As future work, this algorithm can be extended to deal with nonlinear decision boundaries using the kernel trick, and tested in a more extensive set of cases, comparing it to other related approaches. The ideas of learning from label proportions could be extended to other classification paradigms, such as monotonic or ordinal classification, where this type of problems also arise.

Acknowledgment

This work was partly financed by a grant provided by the TIN2014-54583-C2-1-R project of the Spanish Ministry of Economy and Competitively (MINECO), by FEDER Funds and by the P11-TIC-7508 project of the Junta de Andalucía, Spain.

References

[1] J. Hernández-González, I. Inza, and J. A. Lozano, "Weak supervision and other non-standard classification problems: a taxonomy," *Pattern Recognition Letters*, vol. 69, pp. 49 – 55, 2016.

[2] R. Jin and Z. Ghahramani, "Learning with multiple labels," in *Advances in neural information processing systems*, 2002, pp. 897–904.

[3] O. Chapelle, B. Schölkopf, and A. Zien, *Semi-Supervised Learning*, 1st ed. The MIT Press, 2010.

[4] J. Hernández-González, I. Inza, and J. A. Lozano, "Learning bayesian network classifiers from label proportions," *Pattern Recognition*, vol. 46, no. 12, pp. 3425 – 3440, 2013.

[5] K. Fan, H. Zhang, S. Yan, L. Wang, W. Zhang, and J. Feng, "Learning a generative classifier from label proportions," *Neurocomputing*, vol. 139, pp. 47 – 55, 2014.

[6] N. Quadrianto, A. J. Smola, T. S. Caetano, and Q. V. Le, "Estimating Labels from Label Proportions," *Journal of Machine Learning Research*, vol. 10, pp. 2349–2374, 2009.

[7] M. Stolpe and K. Morik, "Learning from label proportions by optimizing cluster model selection." in *ECML/PKDD (3)*, ser. Lecture Notes in Computer Science, D. Gunopulos, T. Hofmann, D. Malerba, and M. Vazirgiannis, Eds., vol. 6913. Springer, 2011, pp. 349–364.

[8] D. R. Musicant, J. M. Christensen, and J. F. Olson, "Supervised learning by training on aggregate outputs," in *Proceedings of the 7th IEEE International Conference on Data Mining (ICDM)*, pp. 252–261.

[9] S. Chen, B. L. 0015, M. Qian, and C. Zhang, "Kernel k-means based framework for aggregate outputs classification." in *ICDM Workshops*, Y. Saygin, J. X. Yu, H. Kargupta, W. W. 0010, S. Ranka, P. S. Yu, and X. Wu, Eds. IEEE Computer Society, 2009, pp. 356–361.

[10] F. X. Yu, D. Liu, S. Kumar, T. Jebara, and S. Chang, "\(\propto\)svm for learning with label proportions," in *Proceedings of the 30th International Conference on Machine Learning, ICML*, 2013, pp. 504–512.

[11] A. Izenman, "Linear discriminant analysis," in *Modern Multivariate Statistical Techniques*, ser. Springer Texts in Statistics. Springer New York, 2008, pp. 237–280.