Learning With Auxiliary Less-Noisy Labels

Yunyan Duan and Ou Wu

***Abstract*— Obtaining a sufficient number of accurate labels to form a training set for learning a classifier can be difficult due to the limited access to reliable label resources. Instead, in real-world applications, less-accurate labels, such as labels from nonexpert labelers, are often used. However, learning with less-accurate labels can lead to serious performance deterioration because of the high noise rate. Although several learning methods (e.g., noise-tolerant classifiers) have been advanced to increase classification performance in the presence of label noise, only a few of them take the noise rate into account and utilize both noisy but easily accessible labels and less-noisy labels, a small amount of which can be obtained with an acceptable added time cost and expense. In this brief, we propose a learning method, in which not only noisy labels but also auxiliary less-noisy labels, which are available in a small portion of the training data, are taken into account. Based on a flipping probability noise model and a logistic regression classifier, this method estimates the noise rate parameters, infers ground-truth labels, and learns the classifier simultaneously in a maximum likelihood manner. The proposed method yields three learning algorithms, which correspond to three prior knowledge states regarding the less-noisy labels. The experiments show that the proposed method is tolerant to label noise, and outperforms classifiers that do not explicitly consider the auxiliary less-noisy labels.**

***Index Terms*— Maximum likelihood approach, noisy degrees, noisy labels, soft constraints.**

I. Introduction

One of the most common assumptions in traditional classification algorithms is that observed labels reflect true classes; otherwise the performance of the classifier can be affected and the complexity of the model is increased [1]. However, this assumption is not met in real-world situations, where the involvement of human labelers naturally introduces label noise, which is often due to the influence of perceptual errors, subjective points of view, and uncertainty resulting from insufficient evidence [2].

To address this issue, noise-tolerant learning algorithms have been introduced in the last decade, which make use of noisy labels with little or no degradation in the performance of classification algorithms [3]–[5]. Typically, in these algorithms, label noise is modeled by a probabilistic model called flipping probability, in which the observed label of a sample is assumed to be flipped from a true label. A likelihood function or a loss function is then defined based on the flipping probability. The resulting parameter estimations are no longer biased by label noise, and the subsequent classification performance is improved.

In this brief, we propose a learning algorithm as an extension of these noise-tolerant approaches. Similar to these methods, we use flipping probability to model label noise; however, our method differs in that, in addition to abundant noisy labels, it also includes auxiliary less-noisy labels for some of the training data. Our work is inspired by recent developments in crowdsourcing learning [6], [7]. In crowdsourcing learning settings, a vast number of noisy labels are provided by a large number of nonexpert labelers. Nowadays, these labels are easy to acquire through network platforms, and they are collected and utilized to infer a not only consentaneous label but also a true one. Despite the results achieved, model performance shows significant improvements when accurate labels are available in even a small part of the training samples [8], [9]. Likewise, in noise-tolerant learning, an expert can also be invited to provide labels for a small part of the training data, within the acceptable time and expense constraints. These auxiliary expert-generated labels need not be perfectly accurate as assumed in previous studies; more realistically and reasonably, they are instead assumed to be less noisy than labels generated by the nonexperts (or, the crowds).

The consideration of these less-noisy labels is also driven by real-world applications, for example, the work of radiologists, which involves determining (i.e., labeling) whether a suspicious region in a medical image is cancerous. Although accurate (noiseless) labels can be obtained through tissue biopsy, this is an expensive and dangerous procedure. As an alternative, radiologists provide labels that are their professional estimates and that are not guaranteed to be noiseless. In practice, the expense of consulting radiologists varies and positively correlates with the expertise level of the individual. A dilemma thus appears when considering budget and resource limitations: junior radiologists require less remuneration but the labels they provide can be noisy, while senior radiologists provide high-quality labels but at a higher cost. The current line of research contributes to resolving this dilemma.

Taking both noisy and less-noisy labels into consideration, the proposed method uses flipping probability to model the noise for both kinds of labels, and then trains classifiers using a maximum-likelihood estimation approach. Specifically, the necessary parameters, including flipping probabilities and coefficients of a logistic regression classifier, are estimated by maximizing the likelihood based on observations. Depending on the extent to which the noise rate of the less-noisy labels is known, which we henceforth call prior knowledge, three cases of the learning process are individually considered and discussed. The three cases are analyzed in a progressive manner, with each case featuring looser constraints on the noise rate of less-noisy labels: in the first case, the accurate value of the noise rate is available; in the second case, a range of the noise rate is available; and in the third case, neither the accurate value nor a range is available. For each case, an expectation–maximization (EM) algorithm [10] is used to infer the classifier.

Our main contributions can be summarized as follows.

1) To the best of our knowledge, this is the first time that both noisy and less-noisy labels have been combined to develop a label noise-tolerant classifier, without requiring that less-noisy labels are perfectly accurate. Three basic cases concerning the prior information for the noise rates of less-noisy labels are considered.
2) Experiments on both synthetic and benchmark data sets show that our method is robust even in the presence of a noise rate that is high and asymmetric between classes.

Manuscript received December 9, 2014; revised January 20, 2016; accepted March 18, 2016. This work was supported by NSFC under Grant 61379098. *(Yunyan Duan and Ou Wu contributed equally to this work.) (Corresponding author: Ou Wu.)*

Y. Duan was with the School of Mathematical Science, Peking University, Beijing 100871 China. She is now with the Department of Linguistics, Northwestern University, Evanston, IL 60201 USA (e-mail: yyduan.pku@gmail.com).

O. Wu is with the National Laboratory of Pattern Recognition, Institute of Automation, Chinese Academy of Sciences, Beijing 100190, China (e-mail: wuou@nlpr.ia.ac.cn).

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TNNLS.2016.2546956

The rest of this brief begins with a discussion of the related work in Section II. The proposed method is then presented in Section III, with the results of empirical studies reported and evaluated in Section IV. This brief ends with a discussion and conclusion in Section V.

II. Related Work

A. Learning With Noisy Labels

Label noise is a complex phenomenon. To narrow down its scope, we focus on noise that originates from a stochastic process where erroneous labels are sample-independent. Noise induced by systematic errors or anomalous outliers is beyond the scope of this brief, though it poses important questions as well [11].

In classification tasks, approaches dealing with label noise can be roughly categorized into three main veins [1]. Some methods are label noise robust, i.e., their performances are not affected by the presence of label noise. Algorithms based on the 0–1 loss function are naturally noise resistant, and strategies that prevent algorithms from overfitting (e.g., regularization) also eliminate the influence of noise [12]. Widely used classifiers, however, are usually not label noise resistant in nature, for instance, support vector machines and AdaBoost [1].

Another approach to dealing with label noise is data cleansing. By detecting samples whose labels are suspected to be corrupt, these samples can be removed [12], reclassified [13], or addressed in a way combining both solutions before data are fed into the training phase [14], [15]. As this approach aims to improve the quality of the training data and has little to do with the complexity of the training algorithms, it can be directly embedded into the classification model. It is possible, however, that some methods may remove too many samples, resulting in a reduction of the model's power [16].

The third approach is modeling the label noise of training data in order to develop noise-tolerant classifiers. These methods usually consist of two components: a noise model and a classification model. Both models play a role in the training phase, with only the classification model being applied in the test phase. Probabilistic models likewise are prevalent in noise modeling. Lawrence and Schölkopf [4] proposed a probabilistic model of a kernel Fisher discriminant using the EM algorithm to update the probability of samples incorrectly labeled. This work further inspired many methods, extending the original method by combining different classifiers and by applying the method in different fields [3], [5], [17]–[19].

Apart from the above approaches, neurofuzzy classifiers [20], [21] are generally used to handle noisy training data (including label noise) due to the advantages offered by adopting fuzzy theory. This approach is fundamentally different from the line followed in our work, however, since neurofuzzy classifiers do not explicitly model the distributions of noisy labels.

B. Learning With Crowd-Generated Labels

With the advent of online annotation networks, crowdsourcing and the analysis of crowdsourcing data have appeared in recent years [7], [8], [22], [23]. Raykar *et al.* [7] presented a model for inferring true labels from labels provided by multiple annotators. They assumed that an observed label depends on both the true label and the reliability of the annotator. They set true labels as a latent variable and used flipping probability to model the annotator's reliability, a technique that was similar to the noise model proposed in [4]. An extension of this work was presented in [8]. In addition to noisy labels from crowdsourcing, they added partial labels from an expert. However, this expert had to be exactly correct, which may be infeasible due to the low accessibility of true labels in real situations. Our method, instead, does not require that the expert labels are exactly accurate.

III. Proposed Method

In the method described here, the training data consists of two parts. The first part contains p independent samples with both noisy and less-noisy labels, $D_p = \{x_i, y_{Li}, \bar{y}_i\}_{i=1}^{p}$, where x_i denotes the d-D feature vector of a sample, $\bar{y}_i \in \{0, 1\}$ denotes a noisy label of the sample, and $y_{Li} \in \{0, 1\}$ denotes a less-noisy label of the same sample (the subscript L indicates less noisy). The second part of the training data contains f independent samples with only noisy labels, $D_f = \{x_j, \bar{y}_j\}_{j=1}^{f}$. As less-noisy labels are available only in a small portion of the data in real situations, p is usually much smaller than f.

Given the training data $D = D_p \cup D_f$, the current method aims to learn a classifier that can be used to classify testing data. Flipping probabilities are adopted to model the relationship between true labels and noisy labels (including less-noisy labels). The parameters of the flipping probability model and the classifier are then estimated through a maximum likelihood approach. Furthermore, using the EM optimization procedure, the process of parameter inference for the model is presented for three different cases, depending on the prior knowledge of the noise rates of the auxiliary less-noisy labels. Our method is referred to as learning with less-noisy data (LLND) for simplicity. First, flipping probability is defined.

A. Flipping Probability

Similar to [4], we introduce a latent variable to represent the ground-truth label and assume that noise is generated by a coin-flipping process. That is, the observable label (either a less-noisy label y_{Li} or a noisy label $\bar{y}_i$) of any sample x_i is generated from the true label ($\hat{y}_i$) through a probabilistic mechanism, and is the same as the true label with some probability pr while flipping to the other side with probability $1 - pr$. Here the flipping probability is assumed to be independent of the sample, i.e., for any sample in the training data, the corresponding probabilities remain the same. Flipping probabilities are defined as follows:

$$\alpha_L := p(y_{Li} = 1 | \hat{y}_i = 1) \tag{1}$$

$$\beta_L := p(y_{Li} = 0 | \hat{y}_i = 0) \tag{2}$$

$$\bar{\alpha} := p(\bar{y}_i = 1 | \hat{y}_i = 1) \tag{3}$$

$$\bar{\beta} := p(\bar{y}_i = 0 | \hat{y}_i = 0). \tag{4}$$

As label y_L is assumed to be less noisy than label $\bar{y}$, we have $\alpha_L > \bar{\alpha}$ and $\beta_L > \bar{\beta}$.

B. Maximum Likelihood Approach

The parameter set $\Omega = \{w, \bar{\alpha}, \bar{\beta}, \alpha_L, \beta_L\}$ is the set of all parameters that need to be estimated with this approach, where w is a d-D classifier coefficient (corresponding to the feature vector x), and the other four parameters are flipping probabilities as defined in (1)–(4). Generally, the estimation of the optimal parameter set $\hat{\Omega}_{\mathrm{ML}}$ uses a maximum likelihood approach that can be represented as follows:

$$\hat{\Omega}_{\mathrm{ML}} = \{\hat{w}, \hat{\bar{\alpha}}, \hat{\bar{\beta}}, \hat{\alpha}_L, \hat{\beta}_L\} = \arg\max_{\Omega}\{\ln[p(D|\Omega)]\}. \tag{5}$$

The classifier used in this brief is assumed to be a logistic regression classifier, in which w is the coefficient

$$p(\hat{y} = 1 | x, w) = \frac{1}{1 + e^{-w^T x}} = \sigma(w^T x). \tag{6}$$

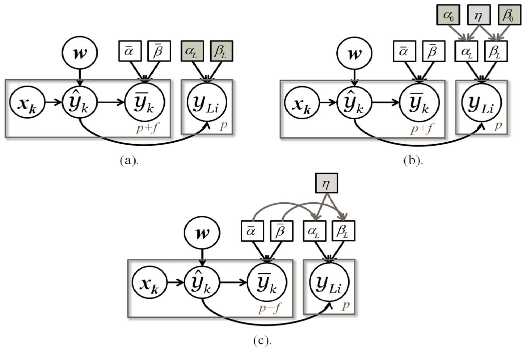

Fig. 1. Graphical representation of observation, ground-truth labels, and the involved parameters. (a) Case 1. (b) Case 2. (c) Case 3.

To estimate $\hat{\Omega}_{\mathrm{ML}}$ through a maximum likelihood approach, the likelihood $p(D|\Omega)$ is calculated as follows:

$$\begin{aligned}
p(D|\Omega) &= \prod_{i=1}^{p} p(x_i, \bar{y}_i, y_{Li}|\Omega) \prod_{j=1}^{f} p(x_j, \bar{y}_j|\Omega) \\
&= \prod_{i=1}^{p} p(\bar{y}_i, y_{Li}|x_i, \Omega) p(x_i) \prod_{j=1}^{f} p(\bar{y}_j|x_j, w, \bar{\alpha}, \bar{\beta}) p(x_j) \\
&\propto \prod_{i=1}^{p} p(\bar{y}_i, y_{Li}|x_i, \Omega) \prod_{j=1}^{f} p(\bar{y}_j|x_j, w, \bar{\alpha}, \bar{\beta}) \\
&= \prod_{i=1}^{p} \sum_{\hat{y}_i \in \{0,1\}} p(\bar{y}_i, y_{Li}|\hat{y}_i, \alpha_L, \beta_L, \bar{\alpha}, \bar{\beta}) p(\hat{y}_i|x_i, w) \\
&\quad \cdot \prod_{j=1}^{f} \sum_{\hat{y}_j \in \{0,1\}} p(\bar{y}_j|\hat{y}_j, \bar{\alpha}, \bar{\beta}) p(\hat{y}_j|x_j, w) \\
&= \prod_{i=1}^{p} \sum_{\hat{y}_i \in \{0,1\}} p(\bar{y}_i|\hat{y}_i, \bar{\alpha}, \bar{\beta}) p(y_{Li}|\hat{y}_i, \alpha_L, \beta_L) p(\hat{y}_i|x_i, w) \\
&\quad \cdot \prod_{j=1}^{f} \sum_{\hat{y}_j \in \{0,1\}} p(\bar{y}_j|\hat{y}_j, \bar{\alpha}, \bar{\beta}) p(\hat{y}_j|x_j, w). \qquad (7)
\end{aligned}$$

Indeed, the definition of (7) does not consider the basic premise in this brief that the label y_L, which is provided by an invited expert, is less noisy than the label $\bar{y}$. Moreover, in practice, we do have some further prior knowledge about the noise rates (i.e., α_L, β_L) of less-noisy labels, rather than being completely agnostic. The reason is that the less-noisy labels are given by an expert with background knowledge (e.g., a professional degree and labeling experience); interaction with this expert can thus provide valuable prior knowledge about the noise rates. Three basic cases are examined and abstracted for the prior knowledge of the noise rates of less-noisy labels. In the first case (Case 1), the exact values of α_L and β_L are known; in the second case (Case 2), the lower bounds of α_L and β_L are known, which are referred to as α_0 and β_0; and in the third case (Case 3), we only know that less-noisy labels maintain better consistency with true labels than with noisy labels, and thus α_L and β_L are larger than $\bar{\alpha}$ and $\bar{\beta}$.

In all the three cases, prior knowledge can be expressed as a set of mathematical constraints. The next section presents the optimization of (5) under the three cases using EM.

C. EM-Based Optimization

The graphical representations of observation D, true labels $(\hat{y}_i)$, and the corresponding parameters (i.e., classifier coefficient and flipping probabilities) under the three cases are shown in Fig. 1. Based on Fig. 1 and the EM procedure, we obtain a concrete learning algorithm for each case respectively, namely, LLND-1, LLND-2, and LLND-3.

Case 1: In this case, α_L and β_L are given. A special setting of this case is when $\alpha_L = \beta_L = 1$. In other words, less-noisy labels are actually true labels. This special setting has been investigated in crowdsourcing learning [8]. In the present model, a more general setting is considered as α_L and β_L are not necessarily equal to 1.

Based on (7) and Fig. 1(a), we obtain

$$\begin{aligned}
\ln p(D|\Omega) &= \sum_{i=1}^{p} \{\hat{y}_i \ln \bar{m}_i m_{Li} P_i + (1-\hat{y}_i) \ln \bar{n}_i n_{Li} (1-P_i)\} \\
&\quad + \sum_{j=1}^{f} \{\hat{y}_j \ln \bar{m}_j P_j + (1-\hat{y}_j) \ln \bar{n}_j (1-P_j)\} \qquad (8)
\end{aligned}$$

where

$$\begin{aligned}
P_i &= \sigma(w^T x_i) && (9) \\
\bar{m}_i &:= p(\bar{y}_i|\hat{y}_i = 1, \bar{\alpha}) && (10) \\
\bar{n}_i &:= p(\bar{y}_i|\hat{y}_i = 0, \bar{\beta}) && (11) \\
m_{Li} &:= p(y_{Li}|\hat{y}_i = 1, \alpha_L) && (12) \\
n_{Li} &:= p(y_{Li}|\hat{y}_i = 0, \beta_L). && (13)
\end{aligned}$$

Since the log-likelihood function contains the latent variable $\hat{y}$, we use the EM algorithm to estimate the parameters.

E-Step: In this process, we calculate the expectation of the log-likelihood [see (8)] of the observed data with respect to the ground-truth labels $\hat{y}$, the observed data D, and the parameters obtained from the previous operation Ω' $(=\{w', \bar{\alpha}', \bar{\beta}', \alpha'_L, \beta'_L\})$

$$\begin{aligned}
&\mathbf{E}\{\ln p(D|\Omega)|\Omega'\} \\
&\quad = \sum_{i=1}^{p} \{\mu_i \ln[\bar{m}_i m_{Li} P_i] + (1-\mu_i) \ln[\bar{n}_i n_{Li} (1-P_i)]\} \\
&\quad\quad + \sum_{j=1}^{f} \{\mu_j \ln[\bar{m}_j P_j] + (1-\mu_j) \ln[\bar{n}_j (1-P_j)]\} \qquad (14)
\end{aligned}$$

where $\mu_i = p(\hat{y}_i = 1|x_i, y_{Li}, \bar{y}_i, \Omega')$, $\mu_j = p(\hat{y}_j = 1|x_j, \bar{y}_j, \Omega')$.

Based on Bayes's theorem

$$\mu_i = \frac{\bar{m}'_i m'_{Li} P'_i}{\bar{m}'_i m'_{Li} P'_i + \bar{n}'_i n'_{Li} (1-P'_i)} \qquad (15)$$

$$\mu_j = \frac{\bar{m}'_j P'_j}{\bar{m}'_j P'_j + \bar{n}'_j (1-P'_j)} \qquad (16)$$

where $\bar{m}'_i$, m'_{Li}, P'_i, $\bar{n}'_i$, $\bar{n}'_i$, n'_{Li}, $\bar{m}'_j$, P'_j, and $\bar{n}'_j$ are calculated by using (9)–(13) based on Ω'.

M-Step: Based on the current estimated μ and the observation D, the model parameters $\bar{\alpha}$, $\bar{\beta}$, and w are updated by maximizing the conditional expectation. By setting the gradient of (14) equal to zero, the following estimates are obtained:

$$\bar{\alpha} = \frac{\sum_{i=1}^{p} \mu_i \bar{y}_i + \sum_{j=1}^{f} \mu_j \bar{y}_j}{\sum_{i=1}^{p} \mu_i + \sum_{j=1}^{f} \mu_j} \qquad (17)$$

$$\bar{\beta} = \frac{\sum_{i=1}^{p} (1-\mu_i)(1-\bar{y}_i) + \sum_{j=1}^{f} (1-\mu_j)(1-\bar{y}_j)}{\sum_{i=1}^{p} (1-\mu_i) + \sum_{j=1}^{f} (1-\mu_j)}. \qquad (18)$$

We do not have a closed-form solution for w due to the nonlinearity of the sigmoid function, which necessitates the use of gradient ascent-based optimization methods. We use the Newton–Raphson method to

Algorithm 1 Steps of LLND-1

```
Input: $D$, $\alpha_L$, $\beta_L$, the initial values of $w$, $\bar{\alpha}$, $\bar{\beta}$
Output: $w$, $\bar{\alpha}$, $\bar{\beta}$
Steps:
1. Initialize $\mu_i = y_{Li}$ and $\mu_j = \bar{y}_j$;
2. Given $\mu_i$ and $\mu_j$, estimate the flipping probabilities $\bar{\alpha}$ and
   $\bar{\beta}$ according to Eqs. (17) and (18) and the logistic regression
   coefficient $w$ according to the iteratively performing Eq. (19) till
   converge.
3. Given the flipping probabilities and logistic regression
   coefficient, update $\mu_i$ and $\mu_j$ using Eqs. (15) and (16).
4. Iterate 2 and 3 till converge.
```

estimate w as demonstrated in the following iterations:

$$w^{(t+1)} = w^{(t)} - \eta H^{-1} g$$

$$g(w) = \left[\sum_{i=1}^{p} [\mu_i - \sigma(w^T x_i)] x_i, \ \sum_{j=1}^{f} [\mu_j - \lambda(w^T x_j)] x_j \right]^T$$

$$H(w) = - \begin{bmatrix} \sum_{i=1}^{p} [\sigma(w^T x_i)][1 - \sigma(w^T x_i)] x_i x_i^T \\ \sum_{j=1}^{f} [\sigma(w^T x_j)][1 - \sigma(w^T x_j)] x_j x_j^T \end{bmatrix}. \tag{19}$$

The two steps (the E- and the M-steps) are iterated until convergence. The log-likelihood increases monotonically after every iteration, which implies convergence to a local maximum. To summarize, the steps of LLND-1 for Case 1 are shown in Algorithm 1.

Case 2: In this case, accurate values of α_L and β_L are unavailable; instead, we know their lower bounds α_0 and β_0 [see Fig. 1(b) for the graphical representation]. The optimization problem using the maximum likelihood approach then becomes

$$\hat{\Omega}_{\text{ML}} = \arg\max_{\Omega} \ \{\ln[p(D|\Omega)]\} \quad \text{s.t. } \alpha_L \geq \alpha_0, \beta_L \geq \beta_0. \tag{20}$$

This equation can be rewritten in the following form:

$$\hat{\Omega}_{\text{ML}} = \arg\max_{\Omega} \ln\{p(D|\Omega) p(\alpha_L|\alpha_0) p(\beta_L|\beta_0)\} \tag{21}$$

where

$$p(\alpha_L|\alpha_0) = \begin{cases} \dfrac{1}{1-\alpha_0} & \alpha_L \in [\alpha_0, 1] \\ 0 & \text{otherwise} \end{cases} \tag{22}$$

and

$$p(\beta_L|\beta_0) = \begin{cases} \dfrac{1}{1-\beta_0} & \beta_L \in [\beta_0, 1] \\ 0 & \text{otherwise.} \end{cases} \tag{23}$$

As both (22) and (23) are not differentiable, new definitions for the relevant variables should be introduced. By considering α_L as an example, we define

$$p_\eta(\alpha_L|\alpha_0) = \frac{1}{\aleph(\eta, \alpha_0)} f_\eta(\alpha_L) = \frac{1}{\aleph(\eta, \alpha_0)} \left(\frac{1}{1 + e^{(-\eta(\alpha_L - \alpha_0))}} \right)$$

where $\aleph(\eta, \alpha_0)$ is a normalized factor to ensure that the integral of $p_\eta(\alpha_L|\alpha_0)$ equals 1 and η ($>= 0$) is a factor that reflects the confidence of the prior. When $\eta = 0$, the confidence of the prior is quite low and the constraint in (20) is invalid, while when $\eta = +\infty$, $p_\eta(\alpha_L|\alpha_0)$ approaches (22). The new definition for β_L follows a similar formulation as α_L. The steps of LLND-2 are presented in the online appendix of this brief.

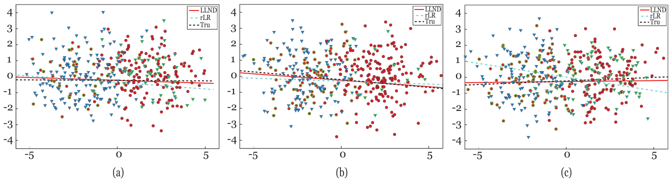

Fig. 2. Comparison of the projection directions learned by LLND and rLR on synthetic data. The mislabeled points are highlighted by green circles. The three label noise types are (a) symmetric low, (b) asymmetric, and (c) symmetric high.

Case 3: In this case, we only know that the less-noisy labels are less noisy than average-quality labels, i.e., $\alpha_L > \bar{\alpha}$ and $\beta_L > \bar{\beta}$. Fig. 1(c) illustrates the graphical representation for this case.

The modeling of this constraint on the noise rates can be directly deduced from Case 2 by merely replacing α_0 and β_0 with $\bar{\alpha}$ and $\bar{\beta}$, respectively. The model estimation and the training procedure follow the maximum-likelihood estimation and EM algorithm. The learning algorithm (LLND-3) for this case is thus similar to LLND-2 with slight modifications.

IV. Experiments and Analyses

We evaluated the proposed LLND algorithm quantitatively on one synthetic data set and three UCI benchmark data sets. Following [5], we artificially added three types of random noise on ground-truth labels to generate the corresponding noisy labels:

1) low symmetric noise, i.e., flipping the labels of 20% of randomly chosen points, which corresponds to $\bar{\alpha} = \bar{\beta} = 0.8$;
2) asymmetric noise, i.e., flipping the labels of randomly chosen points with different noise rates for different classes, where $\bar{\alpha} = 0.7, \bar{\beta} = 0.9$;
3) high noise, i.e., flipping the class labels of 40% of randomly chosen points, with $\bar{\alpha} = \bar{\beta} = 0.6$.

As for the auxiliary less-noisy labels, we set $\alpha_L = \beta_L = 0.95$. The value of η was set to 50.

As learning with auxiliary less-noisy labels has not been explored in the previous literature, we compared the proposed approach LLND with the classical learning-with-noisy-labels method, robust logistic regression (rLR) [3], [17]. Both models were trained on the mixed training data ($D_p \cup D_f$). In the experiments, the updated version of rLR [3] was used which requires the exact values of the flipping probabilities (on the mixed data set $D_p \cup D_f$) as input. In addition, other standard logistic regression classifiers that were trained with three different types of labels were also implemented. The respective labels in the training set consisted of: 1) average-quality noisy labels (Avg); 2) less-noisy labels, which were available in a small portion of the data (Lsn); and 3) true labels (Tru), which were included to train a logistic regression classifier with the highest accuracy possible. We conducted 50 independent runs on each data set and reported the average accuracy as the evaluation criterion.

A. Evaluation on Synthetic Data Sets

We first constructed a synthetic data set by sampling 200 points from two 2-D unit normal distributions centered at (2, 0) and (−2, 0), using 100 points from each distribution. Fig. 2(a)–(c) shows the projection directions learned by the aforementioned methods, respectively, under all the three cases. In all the three cases, compared with the rLR method, the estimation given by LLND was consistently closer to the Tru method (classifier trained by ground-truth labels), which indicated a better noise-tolerance capability of LLND.

TABLE I

PERFORMANCES OF DIFFERENT ALGORITHMS ON UCI BENCHMARK DATA SETS. RESULTS OF LLND ARE PRESENTED UNDER DIFFERENT CASES, RESPECTIVELY (FROM CASE 1 TO CASE 3), AND ARE MARKED IN BOLD IF THEY ARE BETTER THAN OTHER COMPETING ALGORITHMS

Dataset *(dim,n+,n-)*	Noise rates	LLND-1	LLND-2	LLND-3	rLR$_1$	rLR$_2$	Avg	Lsn	NFC	$\tilde{l}_{\log}$	Tru
Thyroid	$\bar{\alpha} = 0.8, \bar{\beta} = 0.8$	87.02	87.98	87.44	86.22	85.17	85.76	80.71	**89.97**	87.80	89.84
(5, 65, 150)	$\bar{\alpha} = 0.7, \bar{\beta} = 0.9$	87.56	87.63	86.80	83.69	82.57	83.48	80.26	**89.04**	80.34	89.74
	$\bar{\alpha} = 0.6, \bar{\beta} = 0.6$	**85.13**	**83.59**	**81.63**	73.81	70.52	73.98	80.23	74.63	83.10	89.57
German	$\bar{\alpha} = 0.8, \bar{\beta} = 0.8$	**72.84**	**74.42**	70.80	72.39	70.43	73.83	68.26	73.28	71.80	76.54
(20, 300, 700)	$\bar{\alpha} = 0.7, \bar{\beta} = 0.9$	**73.00**	**73.70**	70.80	72.67	67.31	68.31	68.35	67.36	71.40	76.55
	$\bar{\alpha} = 0.6, \bar{\beta} = 0.6$	**70.83**	**69.86**	**69.02**	**69.03**	61.79	63.42	68.56	63.02	67.19	76.26
Heart	$\bar{\alpha} = 0.8, \bar{\beta} = 0.8$	76.24	77.44	77.40	77.07	76.80	78.72	66.74	76.28	**82.96**	83.2
(13, 120, 150)	$\bar{\alpha} = 0.7, \bar{\beta} = 0.9$	74.14	78.42	78.33	77.27	76.26	77.99	66.24	74.69	**84.44**	83.44
	$\bar{\alpha} = 0.6, \bar{\beta} = 0.6$	**70.51**	**72.81**	**71.93**	62.33	61.67	64.61	64.15	63.48	57.04	82.89

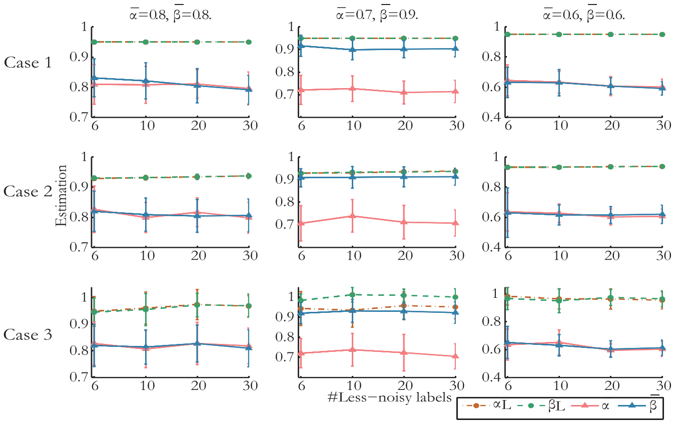

Fig. 3. Effect of less-noisy label amount on the estimation of flipping probabilities.

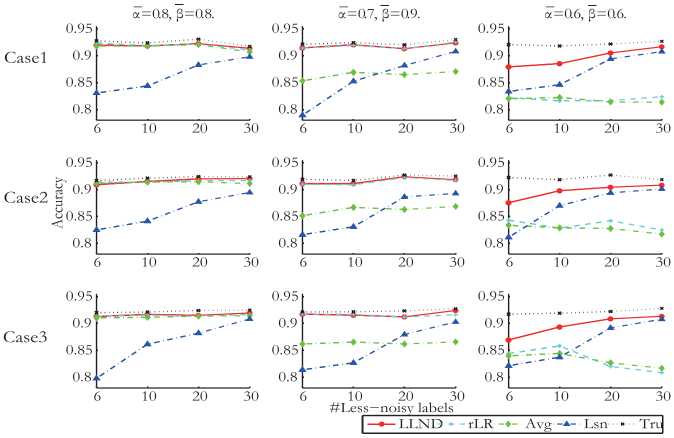

Fig. 4. Effect of the variation of the less-noisy label amount on the performance with different noise rates.

1) Estimation of Flipping Probabilities: Fig. 3 shows the estimated flipping probabilities as a function of the proportion of less-noisy labels. In all the three cases, the estimated flipping probabilities were close to the true noise rates. At the very start, estimations were slightly higher than the true noise rates, but then they quickly approximated the asymptotic line (roughly when the number of less-noisy labels reached 10). Compared with Case 1, the estimation in Case 3 showed a larger variance, showing a decreasing trend due to the prior information in Case 1 being more exact than in Case 3.

2) Model Performance: Fig. 4 shows the accuracies of the classifiers as a function of the proportion of less-noisy labels. As shown in Fig. 4, LLND maintained a high and stable performance in most cases. When labels contained low noise, i.e., $\bar{\alpha} = \bar{\beta} = 0.8$, all algorithms except Lsn performed well and reached nearly the accuracy of the classifier trained by true labels (Tru). The Lsn method, however, was vulnerable to the small size of training samples and showed a clearly rising trend in performance as the proportion of less-noisy labels increased. When the noise was asymmetric, i.e., $\bar{\alpha} = 0.7, \bar{\beta} = 0.9$, only the performances of LLND and rLR were comparable to the performance of Tru. It was quite clear that the Avg was negatively influenced by this kind of noise, which is consistent with previous results [3]. When the noise rate was high, i.e., $\bar{\alpha} = \bar{\beta} = 0.6$, LLND outperformed all other methods. LLND showed a mildly increasing trend as the proportion of less-noisy labels increased, and reached a high performance in a relatively small proportion (compared with Lsn). In particular, both rLR and Avg were significantly influenced by high noise, while LLND was more robust in this setting.

B. Results on Real-World Data Sets

Three data sets from the UCI database were used: German, Heart, and Thyroid. The corresponding results are shown in Table I. In addition, we included the results of a weighted logistic regression classifier, $\tilde{l}_{\log}$, as used in [5], for purposes of comparison. Our experimental settings, which included the data sets and noise rates, were the same as those used in [5]. Two more competing methods were used on these UCI data sets. One is the neuron-fuzzy classifier [20], [21]; the other is an early version of rLR [17], which does not require the exact values of the flipping probabilities as input. Therefore, the updated version of rLR, which requires the flipping probabilities, is denoted by rLR$_1$ and the early version is denoted by rLR$_2$.

From the results in Table I, LLND was superior to rLR (including both rLR$_1$[1] and rLR$_2$) and standard logistic regression with limited labels (i.e., Avg and Lsn), which was consistent under all the three cases. The comparison suggests that the simple mixing of two sets of labels is inferior to explicitly modeling the noisy and less-noisy ones separately. Second, by comparing LLND for

[1] The results showed that LLND-3 significantly outperformed rLR$_1$ on Thyroid at all noise levels, $ps < 0.05$. On Heart, LLND-3 was significantly better at two levels, $ps < 0.05$, while when $\bar{\alpha} = \bar{\beta} = 0.8$, their difference did not reach significance, $p > 0.1$. On German, the results were mixed: LLND-3 and rLR$_1$ did not significantly differ when $\bar{\alpha} = \bar{\beta} = 0.8$; rLR$_1$ outperformed LLND-3 when $\bar{\alpha} = 0.7$ and $\bar{\beta} = 0.9$; LLND-3 outperformed rLR$_1$ when $\bar{\alpha} = \bar{\beta} = 0.6$. To summarize, LLND-3 outperformed rLR$_1$ in six of the nine comparisons, while the two methods performed closely in two comparisons, and only in one comparison did rLR$_1$ perform better. With these results, LLND-3 does achieve a higher accuracy than rLR$_1$ on these benchmark data sets, especially when noise rates were relatively high.

the three cases, a decreasing trend from Case 1 to Case 3 was observed, which demonstrated that our method could benefit from the elaboration of prior knowledge about noise rates. Third, both LLND and rLR_1 were robust to symmetric and asymmetric noise, in contrast to the low performance of Avg in the presence of asymmetric noise. rLR_2 was inferior to rLR_1 on all the three data sets. Fourth, NFC achieved good results on the three data sets when the noise rates were low. However, when the noise rates were increased, its performance decreased significantly. Finally, all methods except Lsn suffered from high noise rates to some extent (compared with the low-noise condition), though LLND showed a relatively smaller loss in accuracy (around 5%) when compared with the 10% loss of Avg and rLR_1. Overall, the proposed LLND methods were competitive and were able to tolerate asymmetric and high noise levels in labels.

V. Discussion and Conclusion

In this brief, we proposed a new method (LLND) that utilized both noisy and auxiliary less-noisy labels to learn a classifier in the presence of label noise. The method was based on the flipping probability of label noise and a logistic regression classifier. By maximizing the likelihood of the observations under an EM framework, the classifier was trained, and the noise rates of noisy labels as well as true labels were estimated jointly. We then implemented the EM-based maximization under three cases according to the extent to which we knew about the noise rate of less-noisy labels. Experiments showed that the proposed method provided accurate estimation for parameters and outperformed noise-tolerant methods proposed in previous studies. When the amount of information on the prior knowledge regarding the noise rates of less-noisy labels was increased, the performance of LLND improved, as is evident from the results witnessed from Case 3 to Case 1. The experimental comparison also indicated that LLND was especially effective in the presence of asymmetric and high noise. As for real-world applications, this research theoretically proved that receiving a small portion of labels from an expert as well as lots of labels from crowds was helpful, as this combination improved the classification accuracy compared with relying entirely on either an expert or the crowds. Future work includes introducing active learning to select samples whose labels are probably noisier than others, and combining the main technical line of neurofuzzy classifiers with the main approach in this brief.

References

[1] B. Frénay and M. Verleysen, "Classification in the presence of label noise: A survey," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 25, no. 5, pp. 845–869, May 2014.

[2] P. Smets, "Imperfect information: Imprecision and uncertainty," in *Uncertainty Management in Information Systems*. Berlin, Germany: Springer-Verlag, 1997, pp. 225–254.

[3] J. Bootkrajang and A. Kabán, "Learning a label-noise robust logistic regression: Analysis and experiments," in *Proc. 14th IDEAL*, 2013, pp. 569–576.

[4] N. D. Lawrence and B. Schölkopf, "Estimating a kernel Fisher discriminant in the presence of label noise," in *Proc. 18th ICML*, 2001, pp. 306–313.

[5] N. Natarajan, I. S. Dhillon, P. K. Ravikumar, and A. Tewari, "Learning with noisy labels," in *Proc. NIPS*, 2013, pp. 1196–1204.

[6] P. G. Ipeirotis, F. Provost, and J. Wang, "Quality management on Amazon Mechanical Turk," in *Proc. HCOMP*, 2010, pp. 64–67.

[7] V. C. Raykar *et al.*, "Learning from crowds," *J. Mach. Learn. Res.*, vol. 11, pp. 1297–1322, Mar. 2010.

[8] H. Kajino, Y. Tsuboi, I. Sato, and H. Kashima, "Learning from crowds and experts," in *Proc. 26th AAAI*, 2012, pp. 107–113.

[9] W. Tang and M. Lease, "Semi-supervised consensus labeling for crowdsourcing," in *Proc. CIR*, 2011, pp. 36–41.

[10] A. P. Dempster, N. M. Laird, and D. B. Rubin, "Maximum likelihood from incomplete data via the EM algorithm," *J. Roy. Statist. Soc. B (Methodol.)*, vol. 39, no. 1, pp. 1–38, 1977.

[11] B. Biggio, B. Nelson, and P. Laskov, "Support vector machines under adversarial label noise," *JMLR*, vol. 20. 2011, pp. 97–112.

[12] C.-M. Teng, "A comparison of noise handling techniques," in *Proc. 14th FLAIRS*, 2001, pp. 269–273.

[13] A. Malossini, E. Blanzieri, and R. T. Ng, "Detecting potential labeling errors in microarrays by data perturbation," *Bioinformatics*, vol. 22, no. 17, pp. 2114–2121, 2006.

[14] S. Fefilatyev *et al.*, "Label-noise reduction with support vector machines," in *Proc. 21st ICPR*, 2012, pp. 3504–3508.

[15] A. L. B. Miranda, L. P. F. Garcia, A. C. P. L. F. Carvalho, and A. C. Lorena, "Use of classification algorithms in noise detection and elimination," in *Proc. 4th HAIS*, 2009, pp. 417–424.

[16] D. R. Wilson and T. R. Martinez, "Reduction techniques for instance-based learning algorithms," *Mach. Learn.*, vol. 38, no. 3, pp. 257–286, 2000.

[17] J. Bootkrajang and A. Kabán, "Label-noise robust logistic regression and its applications," in *Proc. ECML PKDD*, 2012, pp. 143–158.

[18] C. Pal, G. Mann, and R. Minerich, "Putting semantic information extraction on the map: Noisy label models for fact extraction," in *Proc. Workshop on Information Integration on the Web at AAAI*, 2007, pp. 80–85.

[19] D. Wang and X. Tan, "Robust distance metric learning in the presence of label noise," in *Proc. 28th AAAI*, 2014, pp. 1321–1327.

[20] K. Subramanian and S. Suresh, "Human action recognition using meta-cognitive neuro-fuzzy inference system," in *Proc. IJCNN*, 2012, pp. 1–8.

[21] K. Subramanian, R. Savitha, S. Suresh, and B. S. Mahanand, "Complex-valued neuro-fuzzy inference system based classifier," in *Proc. 3rd SEMCCO*, 2012, pp. 348–355.

[22] O. Dekel and O. Shamir, "Vox populi: Collecting high-quality labels from a crowd," in *Proc. COLT*, 2009, pp. 1–10.

[23] Y. Yan *et al.*, "Modeling annotator expertise: Learning when everybody knows a bit of something," in *Proc. 13th AISTATS*, vol. 9. 2010, pp. 932–939.