Whatever you know, just tell me something: Crowd learning with free supervision

Jerónimo Hernández-González[1], Iñaki Inza[1], and Jose A. Lozano[1,2]

[1] Intelligent Systems Group, University of the Basque Country UPV/EHU
P. Manuel de Lardizabal 1, 20018 Donostia, Spain
{jeronimo.hernandez,inaki.inza,ja.lozano}@ehu.eus
[2] Basque Center for Applied Mathematics BCAM
Al. Mazarredo 14, 48009 Bilbao, Spain

Abstract. In the learning from crowds paradigm, many annotators are asked to provide an annotation for the examples of a dataset. Assuming that the participating annotators can be novices in the commended task and, therefore, their annotations can be noisy, different strategies have been developed to build up a reliable labeling from this type of supervision. However, if we consider the possibility of dealing with novice annotators from the beginning, we could assign them a less demanding task. This paper shows that, asking annotators for *certain* information of supervision —whichever they are able to provide according to their knowledge about the commended task— instead of asking them for the complete supervision —a single class label per example—, a reliable labeling can also be obtained and the ultimate objective, that of building a well performing classifier, can be fulfilled.

Keywords: Learning from crowds, weak supervised classification

1 Introduction

An important stage that conditions any machine learning or data mining process is the collection of data. Many different conditions over the data (iid, completeness, etc.) are required for the different techniques to properly work. In the specific case of supervised classification, the *class* variable receives special attention as its information is essential to build a classifier —a function that maps examples to class values. Despite its relevance, gathering the class information for every example in the collected dataset is not always possible. Attempts to learn classifiers using partially labeled examples have given rise to a new subfield, partially or *weakly supervised classification* [7, 12]. Different learning techniques have been developed to efficiently deal with specific lacks of class information, motivated by black-box natural processes, costly labeling collection, etc.

As obtaining reliable labels from a problem expert is usually hard and costly, different approaches have been proposed to rely less on the expert labeling. In this way, the learning from crowds [17] paradigm has recently received much attention in the machine learning community. In this popular paradigm, multiple cheap annotators repeatedly label the examples of a dataset. Noisy labels are expected from these cheap (novice) annotators. Previous key works already showed that learning from noisy labels is possible whenever the mistakes are missing at random [14]. If this is not the case, using multiple annotations with varied backgrounds would guarantee learning well-performing classifiers. This is, in fact, the fundamental basis which the crowd learning paradigm relies on.

In spite of having been almost the only model of (weak) supervision considered so far for labelers' annotations, learning from noisy labels is just another model in the context of weakly supervised classification [12]. Considering that there is no reason to restrict ourselves to a single model of supervision, we propose to learn from the annotations of a set of labelers which can be provided by means of *any* type of class information. That is, the type of supervision that has to be provided is not restricted at all: Each annotator can give a different type of information for different (subsets of) examples.

The hypothesis of this study is that learning well performing classifiers from a set of annotators which provide only partial information is possible. Current learning from crowds approaches are based on the idea that the combination of multiple annotators' opinions compensates their individual (common) errors. Similarly, we rely on the idea that multiple incomplete informations of supervision combine to build up a complete and reliable labeling. It is expected that, if the annotator is not forced to provide a label for each example but *some* class information, the provided information of supervision could be weaker but more reliable. Conceived as a less demanding task, it could be also expected a reduction in the time employed by annotators to complete their tasks.

This paper is structured as follows. In the next section, we position our work with respect to previous related studies of crowd and weakly supervised learning. Then, we provide some insights into the ability to learn from multiple weak labelings in a shortened theoretical framework. In Section 4, the paradigm is formulated and the proposed technique described. Next, the results of our illustrative experimental setting are presented and discussed. We finish drawing a set of conclusions and envisaging possible future works.

2 Related work

Learning from weakly supervised data is an expanding subfield of machine learning which has recently received much attention by the research community [7, 12]. It groups all those problems that aim to learn classification models from a training dataset which is not fully/reliably supervised. Learning from partial or multiple labels [5, 13], learning from noisy labels [21], classical semi-supervision [3, 20], positive-unlabeled [2] or label and positive-unlabeled proportions [16, 10, 11] are a few machine learning problems that have dealt with training datasets characterized by different types of weakly labeled examples.

Learning from noisy labels is a type of weak supervision, as well as the multiple noisy labels provided by a set of non-expert annotators, that is, the learning from crowds paradigm [17]. The model of supervision associated to the annotators of a crowd has been traditionally considered the *noisy labels* model, where each example is individually labeled (by each annotator) although the provided labels are not fully reliable. Many works have shown that it is possible to learn from such a crowd whenever the annotators are not malicious and their errors are not correlated.

We have not been able to find in the related literature a single machine learning framework that considers a crowd of annotators which provides an alternative type of supervision. Restricted by the lack of a class, researches working with the learning from crowds paradigm in clustering domains [19] have considered different types of supervision such a pairwise class constraints. Although ours is a proposal for classification domains, the underlying hypothesis is in line with that of these clustering works: a crowd of annotators can provide useful class information by means of other types of supervision different from the classical noisy individual labels. However, our approach goes beyond as no constraint holds on the type of information the annotator should provide.

3 A theoretical insight: multiple label proportions

The standard learning from crowds, which uses multiple noisy annotations, mainly relies on the strengths of the majority voting strategy [18]. That is, the probability of a crowd of (not malicious) annotators failing at the same time is inversely proportional to the number of annotators. However, considering different models of supervision for the annotators could lead to new techniques not exclusively based on the majority voting strategy.

Let us consider a binary-class problem, with equally probable classes, and a set of m unlabeled examples. Let us assume that two different annotators have access to randomly selected subsets (of size m_i) of the dataset and weakly label the groups with the proportion of examples which belong to each class label (model of supervision: label proportions [10]).

Given a group of examples (bag) and the label proportions annotated by an annotator, the number of consistent completions or possible labelings in this binary scenario is:

$$nCC = \binom{m_i}{m_{ic}} = \frac{m_i!}{m_{i0}! \cdot m_{i1}!}$$

where the label counts, m_{ic}, indicate the number of examples belonging to each class label c in bag B_i. The probability of building up a bag where all the examples belong to the same class ($m_{i0} = 0$ or $m_{i1} = 0$, which implies $nCC = 1$) by random selection is:

$$p(m_{i0} = 0 \vee m_{i1} = 0) = \frac{\binom{m_i}{0} + \binom{m_i}{m_i}}{\sum_{j=0}^{m_i} \binom{m_i}{j}}$$

Similarly, the probability of randomly building up a bag with at most k examples of the class label c is:

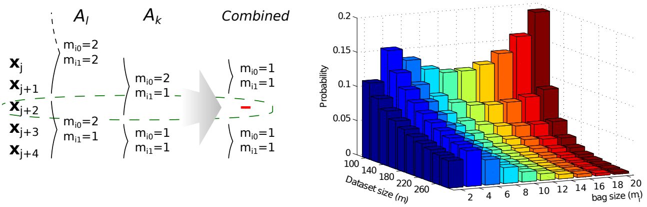

Fig. 1. In left figure, representation of two *incompatible* bags, what reduces the number of consistent completions. As a result, there exists an only possible label assignment for example x_{j+2}. In right figure, probability of randomly producing a pair of incompatible bags as the bag (m_i) and dataset (m) sizes increase (number of annotators fixed to 6).

$$p(m_{ic} \leq k) = \frac{\sum_{j=0}^{k} \binom{m_i}{j}}{\sum_{j=0}^{m_i} \binom{m_i}{j}}$$

Note that the division in groups is not the same for the different annotators. The probability of h examples being provided in the same group to both annotators is:

$$p(|B_a \cap B_b| = h) = \frac{\binom{m_i}{h} \cdot \binom{m-m_i}{m_i-h}}{\binom{m}{m_i}} \tag{1}$$

It can be shown that, given two intersecting bags ($B_a^l \cap B_b^k = I$ with $|I| = h > 0$) labeled by different annotators (A_l and A_k), the number of consistent completions of the examples of the union set, $B_a \cup B_b$, is additionally reduced whenever their respective label proportions are different ($m_{ic}^a \neq m_{ic}^b, \forall c$) and at least one of counts m_{ic}^b is smaller than the intersection set size, $\exists c, b : m_{ic}^b < h$. The probability of randomly generating that such an arrangement is:

$$2 \cdot \sum_{k=0}^{h-1} \left[\frac{\sum_{j=0}^{k} \binom{m_i}{j}}{\sum_{j=0}^{m_i} \binom{m_i}{j}} \cdot \frac{\sum_{j=k+1}^{m_i} \binom{m_i}{j}}{\sum_{j=0}^{m_i} \binom{m_i}{j}} \right] \tag{2}$$

This specific type of arrangement makes some of the consistent completions unfeasible. That is, the more the number of this arrangements in the annotated dataset, the lower the uncertainty about its labeling. Figure 1 displays a graphical example of such a pair of *incompatible* bags, which produce a reduction of the number of possible labelings. In a completely randomly generated environment, the probability of observing a pair of incompatible bags (Eq. 1 × Eq. 2, for all the possible values of h) depends on the bag size and the size of the dataset. It can be observed in Fig. 1 that the probability steadily decreases as the size of the dataset (specifically, the rate m/m_i) grows. Given a fixed dataset size, the probability of finding a pair of incompatible bags grows with large bag sizes (the probability of intersecting bags is larger). However, it is also observed a local maximum in the probability with $m_i = 4$, denoting a local optimum configuration for the probability of generating incompatible bags (Eq. 2) with respect to the probability of generating intersecting bags (Eq. 1).

4 Framework formulation

In the learning from crowds framework, a training dataset composed by m examples $D = \{x_1, x_2, \ldots, x_m\}$ is provided without their ground truth class information. Additionally, a set of labelers annotates the dataset according to their subjective (non-expert) opinion. Each annotator A_l provides *certain* information of supervision for different examples. As in standard supervised classification, the main objective remains learning from the available data well performing classifiers that, given a new unlabeled example, predict its class label.

Although no restriction is imposed on the type of information that labelers provide, there exist two representations that would cover the different types of supervision reported in the related literature [12]: information of supervision individually provided for different examples and information provided globally for groups of examples or *bags*. Any individual information of supervision can be codified by means of a probability distribution over the class labels given the example. Any global information of supervision can be codified by means of a set of consistent completions, that is, each candidate global labeling that jointly assigns label to all the examples of the bag. Thus, each annotator's labeling can consist of a collection of individual and group class informations that weakly supervise different portions of the dataset.

4.1 Methodology

The implemented methodology is based on the Expectation-Maximization (EM) strategy [6], which allows us to learn naive Bayes classifiers, the selected models, in the presence of missing data.

Naive Bayes (NB) classifiers are a type of Bayesian network classifiers [1] which assume conditional independence between the predictive variables given the class variable. By means of this assumption, a simple classifier of fixed structure and relatively small number of parameters can be defined, simplifying to a large extent the learning process. Despite its simple structure, the use of naive Bayes classifiers has reported competitive results in many domains [8]. Its classification rule is:

$$\hat{c} = \operatorname*{argmax}_c p(C = c) \prod_{v=1}^{n} p(X_v = x_v | C = c)$$

where the class variable C is the only parent of all the predictive variables (X_v, with $1 \leq v \leq n$). In general, Bayesian models can be estimated from a set of examples in the case of complete data. Model parameters can be estimated with maximum likelihood estimates by means of frequency counts [9]. The graph of conditional (in)dependencies can also be inferred from data, although it is NP-hard in the general case [4]. The specific learning process of naive Bayes classifiers only involves the estimation of the $p(x_i|c)$ and $p(c)$ probabilities, as no structural learning is required due to its fixed structure.

Different techniques have been proposed in the literature that allow BN models to be learnt in the presence of missing data. Our proposal is based on the EM strategy [6], a widely used and theoretically-founded iterative strategy that, under fairly general conditions, produces a steady increase of the likelihood which converges to a stationary value (commonly, a local maximum) [15]. Each iteration consists of an expectation (E) step, where the missing data is estimated as the conditional expectation of the likelihood given the current fit of the model, and a maximization (M) step, where the model parameters are re-estimated such that the likelihood is maximized given the data completed in the E-step.

Our implementation redefines the standard E-step to calculate the most probable labeling for the training examples. For each annotator, a candidate labeling is generated based on their annotations and the current fit of the model. As aforementioned, two possible general types of supervision are considered. For examples individually annotated, the product of both available probability distributions of the class labels given the example (annotation and model probabilistic prediction) is calculated and normalized to 1 in order to generate the probabilistic labeling. For bags of examples, the joint probability of the different consistent completions is calculated as:

$$p(cc|B_b) = \prod_{j=1}^{m_i} p(C = cc_j | X_1 = x_{j1}, \ldots, X_n = x_{jn})$$

The resulting labeling assigns each example $x_j \in B_b$ to each class label c with probability equal to the normalized addition of the joint probabilities of all the consistent completions which assign class label c to example x_j.

Finally, weighted average is used to combine the different candidate labelings estimated from the annotations of each labeler. This simple combination strategy, a weighted version of the popular majority voting strategy, efficiently aggregates the knowledge of the different annotators to build a common labeling under fair conditions [18]. Moreover, this combination strategy allows our method to deal with inconsistent scenarios where the class information provided by different annotators is contradictory. The combined labeling is used to complete the dataset and re-estimate the model parameters (M-step).

5 Experiments

This set of experiments aims to show the ability to learn, using our basic technique, competitive classifiers from a set of annotators who provide different types

	Dataset size (m)			Dataset size (m)			Dataset size (m)		
	120	180	240	120	180	240	120	180	240
$t = 2$	0.84 ± 0.09	0.89 ± 0.05	0.90 ± 0.05	0.73 ± 0.07	0.74 ± 0.10	0.74 ± 0.08	0.66 ± 0.09	0.68 ± 0.08	0.70 ± 0.08
	0.72 ± 0.14	0.80 ± 0.14	0.82 ± 0.10	0.60 ± 0.09	0.56 ± 0.13	0.61 ± 0.10	0.50 ± 0.09	0.51 ± 0.12	0.56 ± 0.09
$t = 4$	0.87 ± 0.07	0.89 ± 0.05	0.88 ± 0.08	0.77 ± 0.08	0.79 ± 0.06	0.82 ± 0.06	0.71 ± 0.08	0.72 ± 0.08	0.74 ± 0.09
	0.77 ± 0.12	0.77 ± 0.12	0.78 ± 0.16	0.60 ± 0.11	0.66 ± 0.11	0.68 ± 0.12	0.57 ± 0.08	0.57 ± 0.08	0.58 ± 0.10
$t = 6$	0.89 ± 0.07	0.89 ± 0.08	0.89 ± 0.07	0.79 ± 0.09	0.80 ± 0.08	0.83 ± 0.08	0.71 ± 0.10	0.71 ± 0.09	0.74 ± 0.10
	0.78 ± 0.13	0.81 ± 0.09	0.80 ± 0.12	0.66 ± 0.12	0.65 ± 0.11	0.71 ± 0.13	0.57 ± 0.10	0.58 ± 0.09	0.60 ± 0.12
	Class cardinality $\|\mathcal{C}\| = 2$			Class cardinality $\|\mathcal{C}\| = 3$			Class cardinality $\|\mathcal{C}\| = 4$		

Table 1. Results of our technique in terms of mean accuracy and F1 (with associate standard deviations) in upper and bottom lines respectively for different experiments with weak noise-free labelings. Each cell shows the result of a different experimental setting: number of annotators $t = \{2, 4, 6\}$ (in rows), dataset size $m = \{120, 180, 240\}$ (in columns) and class cardinality $|\mathcal{C}| = \{2, 3, 4\}$ (vertical divisions).

of supervision. They also show the behavior of our method when the obtained weak labelings are noisy too.

Synthetic data generation. Generative models are randomly generated using 2DB structures involving 10 binary predictive variables and one class variable. The model parameters have been obtained randomly by sampling a Dirichlet distribution with all the hyper-parameters equal to 1. A set of labeled examples is sampled from the generative model and transformed into a crowd annotated dataset. In order to do so, real labels are removed firstly. For each annotator, labels are randomly changed to simulate noise. Then, groups of examples are randomly selected and their respective (already noised) labels are transformed into weak labels. A probability distribution is generated for individually labeled examples [13]. Label proportions are calculated for bags, allowing for some missing labels ($\sum_{c \in \mathcal{C}} m_{ic} \leq m_i$) [12]. Also two types of mutual label constraints are used: equal or different labels (all the examples of a group show the same –unknown– label or all of them have a different label). The type of weak supervision simulated with each group of examples is randomly selected.

Experimental settings. Each experiment configuration has been repeated 30 times. Using the generative model obtained each time, 10 different datasets are sampled. Our technique is evaluated using a reserved subset of examples (30%). The remaining (70%) examples compose the dataset which is transformed into a dataset with multiple *free* supervisions. Up to 30 different random divisions (70/30) of the sample dataset are tested. Thus, each result shown in this section is the average value obtained from $9,000$ ($30 \times 10 \times 30$) executions of the same experimental configuration.

Thee sample sizes (120, 180 and 240) have been considered, as well as three class variables of different cardinality ($2 - 4$). Two, four and six annotators have been tested, with noise rates (probability of providing a mistaken labeling) ranging from 0.6 to 1.0 (noise-free).

	Noise rate (r)			Noise rate (r)			Noise rate (r)		
	0.60	0.75	0.90	0.60	0.75	0.90	0.60	0.75	0.90
$t = 2$	0.52 ± 0.05	0.61 ± 0.07	0.70 ± 0.07	0.53 ± 0.06	0.60 ± 0.08	0.72 ± 0.09	0.54 ± 0.08	0.65 ± 0.09	0.72 ± 0.10
	0.42 ± 0.08	0.50 ± 0.09	0.57 ± 0.09	0.45 ± 0.10	0.49 ± 0.09	0.58 ± 0.13	0.47 ± 0.09	0.57 ± 0.09	0.61 ± 0.13
$t = 4$	0.56 ± 0.08	0.67 ± 0.07	0.75 ± 0.09	0.63 ± 0.08	0.70 ± 0.10	0.77 ± 0.09	0.60 ± 0.09	0.72 ± 0.10	0.81 ± 0.10
	0.45 ± 0.08	0.56 ± 0.11	0.59 ± 0.11	0.54 ± 0.11	0.58 ± 0.10	0.61 ± 0.11	0.50 ± 0.11	0.60 ± 0.12	0.67 ± 0.15
$t = 6$	0.61 ± 0.08	0.72 ± 0.09	0.80 ± 0.07	0.64 ± 0.09	0.73 ± 0.09	0.78 ± 0.09	0.67 ± 0.11	0.72 ± 0.11	0.77 ± 0.10
	0.50 ± 0.11	0.56 ± 0.10	0.69 ± 0.10	0.53 ± 0.12	0.58 ± 0.10	0.67 ± 0.10	0.59 ± 0.15	0.57 ± 0.11	0.69 ± 0.12
	Dataset size $m = 120$			Dataset size $m = 180$			Dataset size $m = 240$		

Table 2. Results of our technique in terms of mean accuracy and F1 (with associate standard deviations) in upper and bottom lines respectively for different experiments with weak and noise free labelings. Each cell shows the result of a different experimental setting: number of annotators $t = \{2, 4, 6\}$ (in rows), noise rate $r = \{0.6, 0.75, 0.9\}$ (in columns) and dataset size $m = \{120, 180, 260\}$ (vertical divisions).

Discussion. Table 1 shows the summary results of a subset of experiments only considering configurations free of noise ($r = 1.0$). The difference in performance (in terms of both mean accuracy and F1) among classifiers learnt in experiments with different class cardinalities can be explained as a larger number of model parameters has to be estimated from the same amount of data. However, the behavior of the learnt classifiers is equivalent whichever the cardinality of the class variable used in the specific experiments. Increasing both the number of annotators and the size of the dataset consistently enhances their performance. In some cases ($\mathcal{C}$=2), whereas the performance gain related to increasing the number of annotators is noticeable, that related to the enlarged dataset size is less obvious. This behavior could be imputed to the ability of our method to rebuild the real data labeling from this kind of information in order to learn from it. Having been able to recover the real labeling, once the learnt classifiers reach the boundary stablished by the Bayes error, a larger number of examples would not improve the performance of the classifiers.

Although it can be fairly expected a reduction in the amount of mistakes that an annotator makes as the information of supervision that they provide becomes weaker [19], claiming it completely free of error could be controversial. Our strategy is general enough to deal with both weak and noisy multiple labelings, as discerned from the experimental results summarized in Table 2. Only experiments with class cardinality $|\mathcal{C}| = 3$ are shown. The largest improvement is associated to the increase of the number of annotators, as a larger size of dataset produces limited performance gains. Interestingly, the behavior of the classifiers when learning from noisy labelings is similar to that of regular learning from crowds [17, 18]. That is, our simple methodology is able to overcome the issue posed by the use of weakly supervised labelings. The only difficulty preventing our method from learning better performing classifiers would be the presence of noise. Whereas multiple labelers were already needed to deal with weakly labelings, as shown in Table 1, the additional simulation of noisy (weak) labelings shows how the number of annotators which are required to reach to the same results is sharply increased. For example, in experiments with datasets of $m = 180$ examples and $|\mathcal{C}| = 3$ class labels, the performance of a classifier learnt from the noise-free labelings provided by 2 annotators is only reached by 6 annotators if their labelings involve a noise rate $r = 0.75$. Although it has not been included in this paper due to time and space restrictions, a study with real labelers and data to test if less noisy labelings could be fairly expected from annotators which are assigned less demanding tasks, as hypothesized in this work, should be carried out in order to be able to compare our approach with the standard learning from crowds paradigm.

6 Conclusions

In this paper, we present a well-performing learning from crowds frameworks where annotators can provide any kind of class information. The results of the experimental settings show that it is possible to efficiently learn from multiple weak labelings relying on the basic weighted majority voting strategy.

The main hypothesis underlying this work is that asking annotators for a complete labeling is too demanding and induces them to fail. A complete study with real data and labelers should be carried out next to test this hypothesis. Specifically, it would be interesting to test how the reliability of the different labelings is affected when our less demanding task is assigned to the annotators. Could it be shown any correlation between reliability and *degree* of supervision? Is it observed any gain in terms of time required for the labeling process?

Our method is a basic approach which relies on the strengths of the majority voting strategy. On the one hand, a more intricate procedure could be designed trying to identify inconsistencies among annotators. Assuming that lack of contradiction implies correctness, the number of possible labelings could be reduced combining the multiple annotations, which eventually could lead to the actual (not an estimation) full labeling of the training dataset. On the other hand, if larger numbers of annotators or bag sizes were considered, solving this task by means of our current exhaustive technique would become unfeasible. An advance design of our methodology should consider approximate estimation techniques such as Gibbs sampling.

Acknowledgments

This work has been partially supported by the Basque Government (IT609-13, Elkartek BID3A), the Spanish Ministry of Economy and Competitiveness (TIN2013-41272-P) and the University-Society Project 15/19 (Basque Government and University of the Basque Country UPV/EHU). Jerónimo Hernández-González is supported by a post-doc grant (UPV/EHU). Jose A. Lozano is also supported by BERC program 2014-2017 (Basque Government) and Severo Ochoa Program SEV-2013-0323 (Spanish Ministry of Economy and Competitiveness).

References

1. Bielza, C., Larrañaga, P.: Discrete Bayesian network classifiers: a survey. ACM Computing Surveys (CSUR) 47(1), 5 (2014)
2. Calvo, B., Larrañaga, P., Lozano, J.A.: Learning Bayesian classifiers from positive and unlabeled examples. Pattern Recognition Letters 28(16), 2375–2384 (2007)
3. Chapelle, O., Schölkopf, B., Zien, A.: Semi-supervised Learning. The MIT Press (2006)
4. Chickering, D.M.: Learning Bayesian networks is np-complete. Learning from Data: Artificial Intelligence and Statistics V (1996)
5. Cour, T., Sapp, B., Taskar, B.: Learning from partial labels. Journal of Machine Learning Research 12, 1501–1536 (2011)
6. Dempster, A.P., Laird, N.M., Rubin, D.B.: Maximum likelihood from incomplete data via the EM algorithm. Journal of the Royal Statistical Society. Series B (Methodological) 39(1), 1–38 (1977)
7. García-García, D., Williamson, R.C.: Degrees of supervision. In: Proceedings of the 25th Annual Conference on Neural Information Processing Systems Workshops (NIPS). pp. 897–904 (2011)
8. Hand, D.J., Yu, K.: Idiot's Bayes—not so stupid after all? International statistical review 69(3), 385–398 (2001)
9. Heckerman, D.: A tutorial on learning with Bayesian networks. Tech. Rep. MSR-TR-95-06, Learning in Graphical Models (1995)
10. Hernández-González, J., Inza, I., Lozano, J.A.: Learning Bayesian network classifiers from label proportions. Pattern Recognition 46(12), 3425–3440 (2013)
11. Hernández-González, J., Inza, I., Lozano, J.A.: Learning from proportions of positive and unlabeled examples. International Journal of Intelligent Systems (2016), in press
12. Hernández-González, J., Inza, I., Lozano, J.A.: Weak supervision and other non-standard classification problems: A taxonomy. Pattern Recognition Letters 69, 49–55 (2016)
13. Jin, R., Ghahramani, Z.: Learning with multiple labels. In: Proceedings of Advances in Neural Information Processing Systems 15 (NIPS). pp. 897–904 (2002)
14. Lugosi, G.: Learning with an unreliable teacher. Pattern Recognition 25(1), 79–87 (1992)
15. McLachlan, G.J., Krishnan, T.: The EM Algorithm and Extensions (Wiley Series in Probability and Statistics). Wiley-Interscience (1997)
16. Quadrianto, N., Smola, A.J., Caetano, T.S., Le, Q.V.: Estimating labels from label proportions. Journal of Machine Learning Research 10, 2349–2374 (2009)
17. 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 11, 1297–1322 (2010)
18. Snow, R., O'Connor, B., Jurafsky, D., Ng, A.Y.: Cheap and fast - but is it good? evaluating non-expert annotations for natural language tasks. In: Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP). pp. 254–263 (2008)
19. Yi, J., Jin, R., Jain, A.K., Jain, S.: Crowdclustering with sparse pairwise labels: A matrix completion approach. In: AAAI Workshop on Human Computation (Vol. 2). pp. 47–53 (2012)
20. Zhu, X., Goldberg, A.B.: Introduction to Semi-Supervised Learning. Synthesis Lectures on Artificial Intelligence and Machine Learning, Morgan & Claypool (2009)
21. Zhu, X., Wu, X., Chen, Q.: Eliminating class noise in large datasets. In: Proceedings of the 20th International Conference Machine Learning (ICML). pp. 920–927 (2003)