On the Study of Crowdsourced Labeled Data and Annotators: Beyond Noisy Labels

Iker Beñaran-Muñoz[1]
Advisors: Jerónimo Hernández-González[2] and Aritz Pérez[1]
[1] Basque Center for Applied Mathematics, Al. Mazarredo 14, Bilbao, Spain
Emails: {ibenaran, aperez}@bcamath.org
[2] University of the Basque Country UPV/EHU, P. Manuel de Lardizabal 1, Donostia, Spain
Email: jeronimo.hernandez@ehu.eus

I. INTRODUCTION

In this PhD project, multi-class classification problems are considered. The standard framework is supervised classification: there is a dataset with instances (x) that belong each one to only one class c from a set of possible class labels Ω_C, where $|\Omega_C| > 2$. The goal is to train a classifier ϕ that is as accurate as possible, based on the instances from a part of the dataset called *training set*. The examples in a training set are hand-labeled by means of expert knowledge. However, due to the rising volume of data and limitations regarding time, expert availability and/or features of the data, other techniques have emerged and their use has become more extended.

Recently, and with the development of new technologies in the information era, the use of crowdsourcing has popularized through multiple web platforms (e.g. Amazon Mechanical Turk or CrowdFlower) that allow to process short tasks by using the workforce of thousands of workers of unknown expertise. In the last years, crowdsourcing has been widely used to solve different kinds of problems, such as text correction [1], text translation [2] or malaria diagnostics [3].

This new paradigm has been welcomed in the machine learning community as a means to collect labels for unlabeled instances in a fast way and at a low cost. In the machine learning context, the workers are referred to as *annotators* or *labelers*. *Crowd labeling* is the process of getting noisy labels for the instances in the training set from a set of various non-expert annotators A. In this sense, an annotator $a \in A$ can be seen as a classifier which provides labels with a certain amount of noise. As the annotators are not guaranteed to be experts, many labels are usually gathered for each example. In the *traditional crowdsourcing scenario*, every annotator is asked to select a *single* label for each instance. This crowd labeling approach is referred to as *full labeling* throughout this document.

Crowd learning consists of learning a classifier from a dataset with crowdsourced labels. This learning task could be roughly separated into two stages: (i) label aggregation (to determine the ground truth label of each instance of the training set) and (ii) model inference (to learn a model using the aggregated labels and standard supervised classification techniques). If the collected labels fulfill certain conditions, crowd learning can be as reliable as learning from a single expert in a traditional supervised classification framework [4], [5].

Most of the approaches to crowd learning, mentioned below, focus on the first stage mentioned above (label aggregation). Probably the most popular label aggregation technique is majority voting (MV), which assigns to each instance the label that most annotators have selected for it. In weighted voting [6], the label selection of each annotator is weighted according to their reliability. As the expertise of the annotators is often unknown, their reliability has to be calculated based on the labels they have provided.

Many aggregation methods that also model the reliability of annotators were derived from the expectation-maximization (EM) strategy [7]. This strategy was first implemented to learn with multiple (expert) annotators by Dawid and Skene [8], and has been widely used since then [9]–[14]. This method computes estimates for the ground truth and at the same time computes maximum likelihood estimates for the parameters that model the reliability of the annotators. It consists of two steps that are iterated until convergence: (i) Expectation (E-step), where the expected values of ground truth values are computed using the current parameter estimates and (ii) Maximization (M-step), where the parameters are updated with the new maximum likelihood estimates given the current expected data. There are some methods [9] that train a classifier throughout this process. In the model inference stage, the most common approach is to use the aggregated labels (which may be a single label or a probability distribution over the class labels for each instance [15], for example) to train a classifier. As can be seen, in the traditional crowd learning scenario, usually only the labels provided by the annotators are used to perform aggregation and model inference, disregarding the explanatory variables. An extensive review of different label aggregation and crowd learning techniques can be found in [16].

II. HYPOTHESIS

The main hypothesis of this PhD project is that the process of the traditional crowd learning framework can be enhanced if extra relevant information, currently easily accessible but commonly disregarded, is efficiently taken into account. Two

issues of the traditional crowd labeling approach have been identified. This PhD project is aimed to provide a solution to both of them. In the following subsections, each of them is analyzed separately.

1) Lack of flexibility of the labeling process: As aforementioned, in full labeling annotators are required to provide a single label. This request may be too strict when an annotator is in doubt between two or more class labels. Forcing them to choose only one label even if they are not sure could lead to wrong answers. The main hypothesis of this research line is that a more flexible approach to crowd labeling can extract more information from the available labelers. In this context, this project will study the *candidate labeling* approach, where annotators are allowed to provide a set of labels L (called *candidate set*) instead of a single label for each instance. In this way, the correct class label is more likely to be selected and the doubts of each annotator can be reflected.

This first proposal of this PhD project [17], already sent to an international journal for revision, is inspired by the subfield of weak supervision [18], which groups different supervised learning problems where the information of supervision is incomplete. This proposal is especially based on the partial or candidate labels [19] problem, which assumes that all the training examples are provided together with a set of labels, with the guarantee that the real label is in that set. This concept is extended to the context of crowd learning and allows annotators to provide as many labels as they want when they are not sure enough to choose a single one. Note that, unlike in the original candidate labels problem, in this case it is not guaranteed that the real label is in the set provided by an annotator. Frameworks where the single-label request is relaxed have already been proposed, such as the works by [20] and [21], where annotators can say how sure they are about their annotations, or other works where annotators are allowed to claim that they do not know the answer [22], [23]. Our idea may be seen as a step forward in the same direction.

In social sciences, a similar problem has been extensively studied under the name of *approval voting* [24]–[26]. Without ground truth, the objective is to identify popular (approved) options. When a single option needs to be selected, aggregation is usually carried out as follows (using machine learning terminology): Given an instance x, the label included in most candidate sets is chosen. The studies carried out in this field are not of our interest since there is not an aim of estimating a ground truth or of learning any model. Moreover, the aggregation step disregards the information that the size of the candidate sets can bring: confident labelers will provide fewer labels than the hesitant labelers. Assuming that self-confidence and expertise of the annotators are related, the contribution of each annotator could be weighted by, for example, giving more importance to the candidate sets that contain fewer labels. This idea is addressed in [17] under the name of *candidate voting*.

In [27], they already provided some evidence that workers answer faster using candidate labeling ("checkbox interface") than using full labeling ("radio button interface"). Our hypothesis is that not only is this method less costly, but that more knowledge can be extracted and hence better results can be obtained than with full labeling. Novel aggregation and learning methods will be developed within the candidate labeling scenario in order to achieve more efficient learning, in the sense that less time and annotators are required to obtain similar results as techniques using full labeling.

2) Lack of use of the explanatory variables: In crowdsourcing scenarios, the descriptive information of the features of the instances, available by definition in every supervised classification problem, is rarely used to enhance the label aggregation process. The sporadic use of this information is mainly devoted to the estimation of the reliability of the annotators [28] or to model the difficulty of the instances within a framework of active learning [5]. However, the explanatory information only takes part in the decisions of the aggregation functions indirectly. Up to our knowledge, no aggregation technique in the related literature uses explicitly this information.

The hypothesis of this second line of reseach is that the use of the explanatory features during label aggregation can enhance the performance of these techniques. Indirectly, this could impact on the cost of the labeling task, as a lower number of labels would be required to get a dataset satisfactorily labeled.

In order to deal with this issue, this research line aims to produce aggregation methods that exploit the concept of vicinity to aggregate the ground truth label of an instance in the context of full labeling. This study is based on the intuition that, when the class distribution evolves smoothly with respect to the instance space, the class information included in the neighborhood of an instance can be exploited to estimate its class distribution. Under this smoothness assumption, a lower number of labels for a part of the examples might be necessary to obtain correctly aggregated labels.

As of today, we are studying the combination of both the labels gathered for the instance at hand and the labels collected for its k nearest neighbors for label aggregation. This idea, named as *k-nearest voting*, can be understood as an extension of the majority voting technique described above that takes into account the features of the instances as well. This method preserves the simplicity of basic strategies such as (weighted) majority voting, and is able to exploit the useful information from the explanatory variables.

III. Objectives

The main goal of this PhD project is to develop methods to learn from the crowds using both candidate labeling and information from the features. The idea is to make the aggregation and learning stages more efficient, in the sense that better results are reached without increasing the number of annotators. The objectives are as follows:

O1 To study the benefits and weaknesses of both (i) the candidate labeling framework and related learning strategies, and (ii) the use of the features of the instances for label aggregation and associate voting schemes.

O2 To create real-world datasets in order to test the candidate labeling strategy and different aggregation schemes.

O3 To develop algorithms to learn classifiers from crowdsourced data with candidate labeling and aggregation schemes that take into account the information of the explanatory features too.

O4 To apply the novel methods to the real problem of skin cancer diagnosis through medical images throughout a collaboration with physicians and researchers of a local hospital.

IV. METHODOLOGY

This PhD project follows the general methodology of studying each problem from a theoretic point of view and then through an empirical analysis. The results obtained with the fulfillment of the objectives will be sent to a journal of the JCR for evaluation (preferably to one of the first quartile). A version with preliminary results might be presented in a conference of the field. The tasks planned for the attainment of the different objectives are detailed below:

O1: To study the benefits and weaknesses of both (i) the candidate labeling framework and related learning strategies, and (ii) the use of the features of the instances for label aggregation and associate voting schemes.

Although the objectives are similar for both research lines, the tasks are different and they are explained below for each one of them.

Tasks planned for part (i):

T1 Literature review focused on crowd learning and weak supervision techniques.

T2 Comparison between the candidate labeling (candidate labeling) and aggregation (k-nearest voting) technique and the traditional approaches (full labeling and majority voting, respectively).

T3 Write a paper that formalizes the candidate labeling framework and compare the results obtained with the new and the traditional techniques under different experimental conditions. *NOTE:* A paper with results about candidate labeling has been sent to the journal *Pattern Recognition Letters*.

Tasks planned for part (ii):

T1 Literature review focused on crowd learning and techniques that take into account the features of the instances.

T2 Development of techniques that make use of the explanatory variables (as of today, k-nearest voting has been considered).

T3 Comparison between the new aggregation techniques that make use of the explanatory variables and the traditional approaches.

T4 Write a paper that formalize the new framework and compare the results obtained with the new and the traditional techniques under different experimental conditions. *NOTE:* A paper with preliminary results of the k-nearest voting has been sent to the conference *CIKM 2018*.

O2: To create real-world datasets in order to test the candidate labeling strategy and different aggregation schemes.

T1 Study the available platforms to obtain annotations (Amazon Mechanical Turk, Crowdflower...).

T2 Make the necessary adaptations in order that the chosen platform works with candidate labeling.

T3 Generate datasets that are appropriate to test the proposed techniques associate to candidate labeling against state-of-the-art approaches.

T4 Publish the newly generated datasets in a journal in order to make them accessible.

O3: To develop algorithms to learn classifiers from both (i) crowdsourced data with candidate labeling and (ii) labels aggregated through schemes that take into account the information of the explanatory features.

As in **O1**, the tasks vary from the part (i) to the part (ii) of this objective. The tasks planned for part (i) are as follows:

T1 Literature review focused on crowd learning techniques, especially the ones related to the EM method.

T2 Extension of existing methods to the candidate labeling framework and development of novel techniques to learn classifiers within that scenario.

T3 Write a paper with preliminary results for evaluation at a conference of the area. *NOTE:* A paper with preliminary results of an EM-based method extended to candidate labeling has been approved for presentation at *CAEPIA 2018*.

T4 Write papers of selected methods developed in T2 for both the candidate labeling and the k-nearest voting frameworks.

Tasks planned for part (ii):

T1 Literature review focused on crowd learning techniques and methods that take into account the explanatory variables of the instances.

T2 Introduction of schemes that take into account explanatory variables into known learning techniques, and/or development of novel techniques to learn classifiers within that framework. Also, combination of these novel techniques (e.g., k-nearest voting) and the candidate labeling scenario.

T3 Write papers of selected methods developed in T2, showing results.

O4: To apply the novel methods to the real problem of skin cancer diagnosis through medical images throughout a collaboration with physicians and researchers of a local hospital.

T1 Obtain (a) dataset(s) of medical images suitable for our approach.

T2 Literature review focused on image classification and pattern recognition.

T3 Apply the methods developed in **O3** to the medical images dataset(s).

T4 Write paper describing the dataset(s) used, the experimentation carried out and the results reached in the previous task.

V. Relevance

This PhD project opens two novel research lines to work with in the context of crowd learning. One is the candidate labeling framework. Despite having been used previously under the name of approval voting by researchers from the area of social sciences, that kind of labeling has barely received attention in the machine learning community. Thus, many new learning methods may be built (apart from the ones developed in this project) within this framework, and the new datasets that are generated could be used by others to develop those methods. By posing a more relaxed request to the annotators, more information is expected to be exploited and, in consequence, more efficient learning may be possible than with full labeling.

The second research line is the idea of introducing the explanatory variables of the instances into the aggregation stage. The k-nearest voting, for example, can be combined with different learning models or be inserted as an intermediate step into an aggregation scheme. If the labels collected for a certain instance are combined with those gathered for its neighbors, extra information is exploited in comparison to the traditional crowd learning framework. In situations where there are instances that are not labeled or have few labels, these kind of techniques may be useful.

As mentioned in the objective **O4**, application of the new methods will be carried out in collaboration with reputable researchers from a local hospital. We expect that the results reached have a positive impact in the resolution of the problem of skin cancer diagnosis through medical images and that the new techniques are used in that kind of diagnosis.

References

[1] M. S. Bernstein, G. Little, R. C. Miller, B. Hartmann, M. S. Ackerman, D. R. Karger, D. Crowell, and K. Panovich, "Soylent: a word processor with a crowd inside," *Communications of the ACM*, vol. 58, no. 8, pp. 85–94, 2015.

[2] J. Corney, A. Lynn, C. Torres, P. Di Maio, W. Regli, G. Forbes, and L. Tobin, "Towards crowdsourcing translation tasks in library cataloguing, a pilot study," in *Digital Ecosystems and Technologies (DEST), 2010 4th IEEE International Conference on*. IEEE, 2010, pp. 572–577.

[3] M. A. Luengo-Oroz, A. Arranz, and J. Frean, "Crowdsourcing malaria parasite quantification: an online game for analyzing images of infected thick blood smears," *Journal of medical Internet research*, vol. 14, no. 6, p. e167, 2012.

[4] R. Snow, B. O'Connor, D. Jurafsky, and A. Y. Ng, "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)*, 2008, pp. 254–263.

[5] V. S. Sheng, F. J. Provost, and P. G. Ipeirotis, "Get another label? improving data quality and data mining using multiple, noisy labelers," in *Proceedings of the 14th ACM International Conference on Knowledge Discovery and Data Mining (SIGKDD)*, 2008, pp. 614–622.

[6] D. R. Karger, S. Oh, and D. Shah, "Iterative learning for reliable crowdsourcing systems," in *NIPS*, 2011, pp. 1953–1961.

[7] A. P. Dempster, N. M. Laird, and D. B. Rubin, "Maximum likelihood from incomplete data via the EM algorithm," *J. Royal Stat. Soc. Series B*, vol. 39, no. 1, pp. 1–38, 1977.

[8] A. P. Dawid and A. M. Skene, "Maximum likelihood estimation of observer error-rates using the EM algorithm," *J. Royal Stat. Soc. Series C*, vol. 28, no. 1, pp. 20–28, 1979.

[9] V. C. Raykar, S. Yu, L. H. Zhao, G. H. Valadez, C. Florin, L. Bogoni, and L. Moy, "Learning from crowds," *J Mach Learn Res*, vol. 11, pp. 1297–1322, 2010.

[10] P. Welinder, S. Branson, S. Belongie, and P. Perona, "The multidimensional wisdom of crowds," in *Proc of NIPS 23*, 2010, pp. 2424–2432.

[11] E. Côme, L. Oukhellou, T. Denoeux, and P. Aknin, "Learning from partially supervised data using mixture models and belief functions," *Pattern recognition*, vol. 42, no. 3, pp. 334–348, 2009.

[12] J. Whitehill, P. Ruvolo, T. Wu, J. Bergsma, and J. R. Movellan, "Whose vote should count more: Optimal integration of labels from labelers of unknown expertise," in *Proc. of NIPS 22*, 2009, pp. 2035–2043.

[13] P. L. López-Cruz, C. Bielza, and P. Larrañaga, "Learning conditional linear gaussian classifiers with probabilistic class labels," in *Conference of the Spanish Association for Artificial Intelligence*. Springer, 2013, pp. 139–148.

[14] Y. Zhang, X. Chen, D. Zhou, and M. I. Jordan, "Spectral methods meet em: A provably optimal algorithm for crowdsourcing," in *Advances in neural information processing systems*, 2014, pp. 1260–1268.

[15] R. Jin and Z. Ghahramani, "Learning with multiple labels," in *Proceedings of Advances in Neural Information Processing Systems 15 (NIPS)*, 2002, pp. 897–904.

[16] J. Zhang, V. S. Sheng, J. Wu, and X. Wu, "Multi-class ground truth inference in crowdsourcing with clustering," *IEEE Trans. Knowl. Data Eng.*, vol. 28, no. 4, pp. 1080–1085, 2016.

[17] I. Beñaran-Muñoz, J. Hernández-González, and A. Pérez, "Weak Labeling for Crowd Learning," *ArXiv e-prints*, 2018.

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

[19] T. Cour, B. Sapp, and B. Taskar, "Learning from partial labels," *Journal of Machine Learning Research*, vol. 12, pp. 1501–1536, 2011.

[20] C. Grady and M. Lease, "Crowdsourcing document relevance assessment with mechanical turk," in *NAACL HLT 2010 workshop*, 2010, pp. 172–179.

[21] P. Smyth, U. M. Fayyad, M. C. Burl, P. Perona, and P. Baldi, "Inferring ground truth from subjective labelling of venus images," in *Proc of NIPS 7*, 1994, pp. 1085–1092.

[22] J. Zhong, K. Tang, and Z.-H. Zhou, "Active learning from crowds with unsure option," in *Proc. of 24th IJCAI*, 2015, pp. 1061–1068.

[23] M. Venanzi, J. Guiver, P. Kohli, and N. R. Jennings, "Time-sensitive bayesian information aggregation for crowdsourcing systems," *J. Artif. Intell. Res.*, vol. 56, pp. 517–545, 2016.

[24] S. J. Brams and P. C. Fishburn, "Approval voting," *Am. Polit. Sci. Rev.*, vol. 72, no. 3, pp. 831–847, 1978.

[25] J.-C. Falmagne and M. Regenwetter, "A random utility model for approval voting," *J. Math. Psychol.*, vol. 40, no. 2, pp. 152–159, 1996.

[26] A. D. Procaccia and N. Shah, "Is approval voting optimal given approval votes?" in *NIPS*, 2015, pp. 1801–1809.

[27] S. O. A. Banerjee and D. Gurari, "Let's agree to disagree: A meta-analysis of disagreement among crowdworkers during visual question answering," in *GroupSight Workshop at AAAI HCOMP*, Quebec City, Canada, 2017.

[28] Y. e. a. Yan, "Modeling annotator expertise: Learning when everybody knows a bit of something," in *Proceedings of the 13th International Conference on Artificial Intelligence and Statistics (AISTATS)*, 2010, pp. 932–939.