On the Evaluation and Selection of Classifier Learning Algorithms with Crowdsourced Data

A. Urkullu[1,*], A. Pérez[b], B. Calvo[a]

[a]Department of Computer Science and Artificial Intelligence, University of the Basque Country UPV/EHU, Paseo de Manuel Lardizabal 1, 20018, San Sebastián, Spain
[b]Department of Data Science, Basque Center for Applied Mathematics, Alameda Mazarredo 14, 48009, Bilbao, Spain

Abstract

In many current problems, the actual class of the instances, the ground truth, is unavailable. Instead, with the intention of learning a model, the labels can be crowdsourced by harvesting them from different annotators. In this work, among those problems we focus on those that are binary classification problems. Specifically, our main objective is to explore the evaluation and selection of models through the quantitative assessment of the goodness of evaluation methods capable of dealing with this kind of context. That is a key task for the selection of evaluation methods capable of performing a sensible model selection. Regarding the evaluation and selection of models in such contexts, we identify three general approaches, each one based on a different interpretation of the nature of the underlying ground truth: deterministic, subjectivist or probabilistic. For the analysis of these three approaches, we propose how to estimate the Area Under the Curve (AUC) of the Receiver Operating Characteristic (ROC) curve within each interpretation, thus deriving three evaluation methods. These methods are compared in extensive experimentation whose empirical results show that the probabilistic method generally overcomes the other two, as a result of which we conclude that it is advisable to use that method when performing the evaluation in such contexts. In further studies, it would be interesting to extend our research to multiclass classification problems.

Keywords: Model selection, evaluation, crowdsourced data, AUC, Kendall-tau

[*]Corresponding author
Email address: ari.urkullu@ehu.eus (A. Urkullu)

1. Introduction

The main target in supervised classification is to build models that accurately predict the class value for new, unseen instances. In order to do so, we need a set of instances for which the ground truth, i.e., the true class value, is known. Those instances allow the training of classifiers using learning algorithms and the assessment of trained classifiers. The assessment of the classifiers for the model selection is usually performed through the estimation of a score (e.g., the accuracy or the AUC, among others) based on the comparison of the true classes of the instances with the predicted classes.

Nevertheless, there are problems for which the ground truth is not available. This unavailability can happen due to different reasons, like, for instance, excessive cost, difficulty or risk of gathering the ground truth. Thus, in these problems, there is no set of labeled instances to perform the learning, evaluation or selection of models. Alternatively, for some of these problems, the labels of multiple annotators per instance can be gathered. These annotations, although probably noisy and biased, serve as an alternative to the ground truth. Examples of this general situation can be found throughout the literature [1–18]. With these annotations, tasks such as the learning, evaluation and model selection may be attempted even if the ground truth is not available. Within this strategy falls the concept of crowdsourced data, in which the data are collected from groups of people.

The task of learning from data with crowdsourced labels is a growing discipline that has received much attention in the last decade [1–3, 6–9, 11–22]. This growth is favored by the fact that the conditions for gathering crowdsourced labels for large unlabeled datasets have improved lately, both technically and economically. Technologies, such as the Internet and online platforms, which take advantage of it, have dramatically changed the conditions in which the crowdsourced data can be harvested, with Amazon Mechanical Turk (https://www.mturk.com), MicroWorkers (https://microworkers.com) and Figure Eight (https://www.figure-eight.com) being some examples of the current online platforms available. Therefore, the gathering of labels by crowdsourcing is currently easy, cheap and fast [1–7, 14, 15, 18, 22–25], making this strategy very profitable in comparison to the other alternatives such as, for instance, the use of expert annotators.

As previously mentioned, crowdsourced labels are used in order to confront the lack of a ground truth, which may be due to different reasons [26]. On the one hand, the unavailability of the true values of the labels may be because of their non-deterministic nature. This is likely to happen in problems where some sort of subjectivity is involved. For instance, this is the case of problems from the area of sentiment analysis (e.g., feelings aroused by articles [19]) and problems such as the development of recommender systems of movies [27] or songs [28], among others. On the other hand, the non-availability of the true values of the labels may be due to the cost or difficulty in gathering them. This happens, for example, in problems related with areas such as computer vision (e.g., image segmentation [29], object detection [10]) or geometrical reasoning [30].

In classification problems, as far as we know, the evaluation and the selection of models from crowdsourced data have not been dealt with explicitly before. Instead, in papers covering the learning from multiple annotators, when evaluation and model selection is performed, it is done through simulation or similar approaches [9, 11–13]. Such occurrence implies a research gap, which supposes the rationale behind our study, since performing an accurate ranking of models even when the ground truth is not available is crucial in order to achieve a sensible model selection. In fact, in actual problems in which the ground truth is not available, and therefore in which the evaluation and selection of models can hardly be done even through simulation, at least two questions arise. The first question consists of whether, for such contexts, a useful evaluation method that takes advantage of the labels issued by the annotators can be proposed. The second question is how well performs the evaluation method proposed for that context when assessing the models. Consequently, in this paper we tackle the problem that can be defined as the evaluation and selection of models in binary classification problems in which the ground truth is unavailable and in which crowdsourced labels are available instead. Briefly, the main objective of our research is to explore the proposal, assessment and comparison of evaluation methods in such problems, focusing on the selection of models, so as to tackle the aforementioned research gap. In order to carry out that exploration, we specifically seek to quantitatively assess the behavior of different evaluation methods capable of dealing with this kind of contexts, in terms of their ability to conduct a sensible model selection. Namely, an outline of which evaluation method is more appropriate for model selection depending on the specific characteristics of the given problem is sought here.

The methodology we have followed to conduct our research is composed of different sequential steps. The first one is the identification of the research gap, which has already been exposed. The rest of them are explained in later sections in the same order they were conducted. Briefly, those remaining steps are the definition of the problem, the proposal of a solution, the evaluation of the capability of the solution to tackle the problem, the derivation of conclusions from the evaluation and the design of future research lines based on those conclusions.

In order to pursue the aforementioned objective, we propose a taxonomy composed of three approaches to tackle the evaluation and model selection in binary classification problems in which the ground truth is unavailable, based on three different views of the underlying ground truth. Now we give a brief description of the essence of each approach, while later in the text (Section 3) how each approach has been identified is explained. First, the deterministic view assumes that every instance has a deterministic label value. Secondly, the subjectivist view assumes that each annotator expresses an alternative deterministic ground truth through the subjectivity of their opinion. Finally, the probabilistic view assumes that the ground truth has a probabilistic nature. These different approaches can be used to adapt the estimators of evaluation measures in different ways. Next, the supervised estimation of the AUC is adapted to the three general approaches presented, since the AUC is used as the evaluation measure of reference. In order to adapt the estimation of the AUC to the probabilistic interpretation of the ground truth, we use a generalization based on the Kendall-Tau distance [31–34]. Finally, in the experimentation, we compare these evaluation methods through simulation, measuring how similar they are to the evaluation performed with the ground truth, in terms of rankings of classifiers.

This paper is structured as follows. To begin with, in Section 2, the problem tackled in this paper is defined and formalized. Next, in Section 3, our proposal consisting of three different general approaches to perform the evaluation and model selection in the problem at hand and three specific evaluation methods, each one belonging to a different approach, is explained. In Section 4, the experimentation is explained while, in Section 5, the results achieved are described. Next, in Section 6 the most important aspects of this work are discussed and the main conclusions are given. Finally, in Section 7 the recommendations derived from this work and the identified further studies to carry out are given.

2. Problem

In this section, we specify the problem that we deal with in this paper. Briefly, in this work we tackle the problem of the selection of classifier learning algorithms in binary classification problems in which the ground truth is unavailable and in which the labels of multiple annotators are available instead. In addition, we limit the scope of the problem by considering that there is no information available regarding the reliability of the annotators. In such a problem, our objective is to explore the proposal, assessment, and comparison of different estimators of a given evaluation measure in their task of performing the selection of classifier learning algorithms under those circumstances.

In order to illustrate the problem, let M^+ and M^- be two models learned from two different classifier learning algorithms applied to a given binary classification problem. In such a context, let e^+ and e^- be the associated errors of M^+ and M^-, respectively, in terms of a given evaluation measure (e.g., the accuracy or the AUC, among other possibilities) that assesses the ability of M^+ and M^- as binary classifiers, verifying that $e^+ < e^-$. Let S be an estimator of the evaluation measure of reference, S requiring for its computation the availability of the ground truth for a given set of instances. Let $S(M^+)$ and $S(M^-)$ be the distributions of the estimations of the errors of M^+ and M^- (e^+ and e^-) according to S. Let S' be another estimator of the evaluation measure of reference that, unlike S, does not require for its computation the availability of the ground truth for a given set of instances, it being capable of using instead the labels of multiple annotators for the given set of instances. Let $S'(M^+)$ and $S'(M^-)$ be the distributions of the estimations of the errors of M^+ and M^- (e^+ and e^-) according to S'. Since what is being sought is the selection of classifier learning algorithms, it is sensible to assess the goodness of a given estimator (S, S', ...) in terms of its probability to issue the same ranking for M^+ and M^- that could be generated with the assessments of M^+ and M^- that the evaluation measure of reference does. Namely, the goodness of S and of S' can be quantified through the probabilities $P(S(M^+) < S(M^-))$ and $P(S'(M^+) < S'(M^-))$, respectively.

Besides, when the problem is generalized in order to rank more than two models, then there are no longer just two rankings, one correct (M^+ better than M^-) and one incorrect (M^+ worse than M^-). Instead, if there are k classifiers, there are k! different possible rankings, among which only one is a totally correct ranking and only one is a totally incorrect ranking, while the rest of the rankings have some degree of correctness. In this more complex case, the goodness of S and of S' can be quantified through the sum of the probabilities of all the different possible rankings weighted accordingly to their similarities with the correct ranking.

Once the general problem has been exposed, we dedicate the next two subsections to giving further details of it. First, in Subsection 2.1 we narrow down the scope of the problem addressed by selecting a specific evaluation measure of reference, the AUC. Secondly, in Subsection 2.2 we narrow down the scope of the problem again, limiting it to the context of crowdsourced data, which is a growing context that is receiving much attention lately due to its aforementioned advantages.

2.1. Evaluation measure of reference

In this subsection, we analyze some of the most popular evaluation measures used in binary classification problems in which the ground truth is available, in order to select one of them to be the evaluation measure of reference with which to carry out our research. In contexts where the ground truth is available, a direct estimation of a given evaluation measure can usually be calculated comparing the outputs of the classifiers with the true class. Among the measures used in these contexts when binary classification is done, some of the most common ones are the accuracy (or, equivalently, the error), the specificity, the sensitivity, different combinations of both and the AUC [9, 12, 13].

The accuracy can be defined as the probability of a randomly chosen instance to be correctly classified. Unfortunately, the accuracy has handicaps, such as its sensitivity to imbalanced proportions between the real classes [35–37] and its dependency on the decision threshold [38] that defines the trade-off between the reliability of the predicted positive and the predicted negative instances.

The sensitivity and specificity are two measures used in binary classification [39],[40] whose meanings are the probability to classify as positive a positive instance and the probability to classify as negative a negative instance, respectively. Those two measures can be combined into a single measure, such as, for instance, the g-means [41], which consists of a geometric average of both measures that has the advantage of being insensitive to unbalanced datasets. Namely, unlike the accuracy, the g-means is invariant regarding the a priori class probabilities. However, since each decision threshold derives specific values for the sensitivity, specificity and g-means, it happens that these metrics are also dependent on the decision threshold [38],[40].

The Area Under the Curve (AUC) of the Receiver Operating Characteristic (ROC) curve, which is frequently referred to simply as AUC, is a measure related with the sensitivity and specificity. In fact, one of the axes of the ROC curve is the sensitivity itself, while the other axis is one minus the specificity. The AUC can be interpreted as the probability of a pair of instances chosen randomly, composed of a negative and a positive instance, to be correctly ordered by the classifier [38, 42]. Compared with the sensitivity and the specificity, the AUC has the advantage of being independent of the decision threshold, as it is based on the pairs of sensitivity and specificity obtained with all the possible thresholds [38]. In addition, the AUC has the advantage of being statistically more discriminative than the accuracy [43, 44].

In conclusion, in this section several measures to perform the evaluation and selection of models in supervised binary classification have been exposed. Considering the relative advantages and disadvantages presented for the different evaluation measures, we choose to use the AUC. Therefore, the problem we deal with in this paper is bounded to the AUC. Namely, in the problem of the selection of models in binary classification problems in which the ground truth is unavailable and in which crowdsourced labels are available instead, we focus on the use of the AUC as the evaluation measure of reference. This implies that the three different estimators proposed in this paper, which are computable in problems in which the ground truth is unavailable, are adaptations of an estimator of the AUC that needs the ground truth to be available for its computation. Specifically, we select the popular estimator of the AUC based on the Wilcoxon or Mann-Whitney statistic [42] to be the estimator of the AUC when the ground truth is available.

2.2. Crowdsourced data context

In crowdsourced contexts, where the ground truth is unavailable, the evaluation is no longer a straightforward task, because the classification cannot be compared with the ground truth to compute a direct estimation of the evaluation measures. Since the only knowledge held about the ground truth is the labels from the annotators, it is worth mentioning some properties related to the crowdsourced data which may affect and hinder the subsequent learning, evaluation and selection of models.

To start with, when the labels are obtained through crowdsourcing, the amount of annotators is usually large [2, 3, 11, 13, 20, 22, 25, 27] and they tend to be non-experts [1–8, 11, 13–15, 19, 23, 24, 35]. Furthermore, it is reasonable to assume that each annotator will label a different number of instances [6, 7, 9, 20, 27]. In addition, the matrix of labels, in which one dimension represents the instances and the other dimension represents the annotators, tends to be sparse. This happens largely because it is unfeasible for each annotator to label a large amount of instances, it being usual for each one to label only a few instances [2, 9, 23, 27]. Finally, differences among the qualities of the annotators are likely to appear [6, 7, 9, 12, 13, 22, 25, 29], it also being possible that they label in an unbalanced way [3, 13, 22, 24].

3. Proposal

In order to expose our proposal, it is convenient to recall that, in the problem at hand, information of the unavailable ground truth is sought through the collection of the labels of multiple annotators per instance in a crowdsourced data context. Since such annotators aim to guess the ground truth through their labels [1–18], we seek to use those labels to perform the evaluation and model selection in the problem tackled. In this regard, we propose three different general strategies that match three interpretations about the underlying ground truth and that can be used to adapt the estimators of different evaluation measures to contexts in which the ground truth is unavailable.

Within each general approach, we propose and describe a particular evaluation method that supposes an adaptation of the selected estimator of the AUC, an estimator that requires the availability of the ground truth for its computation (an estimator like S). As the problem at hand requires, those three adaptations share the particularity of being computable in binary classification problems in which, instead of the unavailable ground truth, crowdsourced labels are available (three estimators like S'). Specifically, the three evaluation methods proposed in this paper have been designed to be as simple and aseptic as possible within each approach, so that the comparison of their performances reflects, to some extent, the different approaches. Specifically, that design is made taking into account that, in this work, the research is focused on problems for which there is no information available regarding the reliabilities of the annotators. In consequence, that simplicity and that asepsis are sought through the attempt to give the same relevance to each piece of information, which consists of a label issued by an annotator. However, because during the experimentation sparse matrices have to be handled and due to the need to deal with instances without any labels, that intention is somewhat hampered.

Therefore, in each of the next three subsections a different general approach is exposed. Besides, within each approach a particular evaluation method is explained as to how to adapt the selected estimator of the AUC to be used when the ground truth is available.

3.1. Deterministic ground truth approach

In this approach, the interpretation is that, for every instance, a deterministic label value exists. This is the case of problems in which, although the ground truth exists in a deterministic form, getting the ground truth is too expensive, risky or difficult. For instance, this situation happens, among other areas, in remote-sensing [10, 45] (e.g., whether or not an object is present in a given picture) and medical diagnosis [9, 46] (e.g., whether or not an individual has a given disease).

Taking into account the aforementioned interpretation, we propose a two-step approach to tackle the evaluation. The first step is to establish a unique estimation of the deterministic ground truth for each instance through a function that uses the labels of the annotators. Afterwards, once the estimation of the deterministic ground truth is established, the second step consists of the use of an ordinary estimator (one that requires the availability of the ground truth) of the evaluation measure.

The first step can be completed, for example, through the majority voting technique, which as mentioned in the literature [9, 12], is a simple and frequently used technique to establish a unique estimation of the deterministic ground truth. Alternatively, when information on the reliability of the labelers is available, a weighted voting can be applied in order to take into account that information.

3.1.1. Deterministic Ground Truth (DGT) method

Within the deterministic ground truth approach, we propose an evaluation method based on the majority voting, i.e., each instance is assigned to the label that most of the labelers issue (solving the ties randomly). We propose the majority voting because it is a straightforward way to estimate the deterministic ground truth while giving the same relevance to each annotator, given that there is no information regarding their reliability. Once the estimation of the deterministic ground truth is computed, the estimation of the AUC is calculated by using the estimator of the AUC based on the Wilcoxon or Mann-Whitney statistic [42].

The essence of this method can be seen in the algorithm of Figure 1, where $\boldsymbol{q}$ represents the outputs of a classifier for a set of m instances and L represents the matrix with the labels of the n annotators for the m instances, in which each row (represented as $L_{i\cdot}$) is related to an instance and each column is related to a labeler. Finally, in Figure 1, let i be the index that denotes the current instance in the loop of the algorithm, let l_i be the outcome of the majority voting of the labelers for instance i, let $\boldsymbol{l}$ be the outcomes of the majority voting of the labelers for all the instances and let *measure* be the outcome of the method.

```
Algorithm 1: DGT
  Input: $\boldsymbol{q}$ and $L$
  Output: measure
  begin
      for $i \leftarrow 1$ to $m$ do
          $l_i \leftarrow$ majority_voting$(L_{i\cdot})$
      end
      measure $\leftarrow$ auc$(\boldsymbol{q}, \boldsymbol{l})$
      return measure
  end
```

Figure 1: Algorithm of the DGT method.

3.2. Subjectivist ground truth approach

In this case, the interpretation is that each annotator has a different subjective view, expressing an alternative deterministic ground truth. This profile is suitable for problems in which, simply, the ground truth does not exist in an objective way. For example, among the problems in which the aforementioned circumstance is present, the assessment of the relevance of books and documents regarding a topic [47], music recommendation [28, 48], recommendation of movies [27], and sentiment analysis [19] can be found.

Considering the exposed interpretation, we outline a two-step approach to deal with the evaluation. The first step is to apply an ordinary estimator (one that requires the availability of the ground truth) of the evaluation measure as many times as the amount of annotators, each time using the alternative deterministic ground truth of a different annotator. In the second step, all the performances are combined somehow through a function in one global value to express a synthesis of the performance.

The combination of the performances of the second step can be achieved, for instance, through the mean or the median. Another option is the use of a weighted average, with weights that take into account the amount of labels issued by each labeler or, when available, the reliability of the labelers, or both.

3.2.1. Subjectivist Ground Truth (SGT) method

In this case, we propose to estimate the AUC values for a given classifier regarding each of the alternative deterministic ground truths defined by each labeler by using the estimator of the AUC based on the Wilcoxon or Mann-Whitney statistic [42]. Next, a weighted average of them is performed to calculate the estimation of the global AUC. Since no information regarding the reliabilities of the annotators is available, by default this method assumes that the reliability of each annotator is the same and therefore their reliabilities do not affect the weights. Specifically, each weight is set to be proportional to the amount of labels that the associated annotator issues. Namely, the weight of a given AUC derived from a given annotator is set to be proportional to the amount of labels issued by that annotator. The reason why this weighting is performed is to give the same relevance to each piece of information, which consists of a label issued by an annotator. The algorithm in Figure 2 summarizes how the method works. Specifically, in Figure 2, let j be the index that denotes the current annotator in the loop of the algorithm, let $L_{\cdot j}$ be the column of matrix L related to the labeler j, let auc_j be the AUC of the given classifier taking the labels of labeler j as the ground truth, and let ***auc*** be the vector of the AUCs of the given classifier taking at each position of the vector the labels of a different annotator as the ground truth.

```
Algorithm 2: SGT
Input: $\boldsymbol{q}$ and $L$
Output: measure
begin
    for $j \leftarrow$ to $n$ do
        $auc_j \leftarrow \mathrm{auc}(\boldsymbol{q}, L_{\cdot j})$
    end
    measure $\leftarrow$ weighted_average($\boldsymbol{auc}$)
    return measure
end
```

Figure 2: Algorithm of the SGT method.

3.3. Probabilistic ground truth approach

In this approach, the interpretation is that the ground truth takes a probabilistic shape. This is the case of problems in which the underlying ground truth exists in a probabilistic form or it is natural to express it in a probabilistic form. For instance, in image processing, some pixels (or voxels, ...) may each show a mix of objects that belong to different classes [49–51]. Therefore, a pixel can naturally have multiple classes, each pixel having a proportion corresponding to the presence of the corresponding class within the pixel. In the case of binary classification, a single continuous value within the range $[0, 1]$ is enough to describe the relative proportions of both classes.

Bearing in mind that interpretation, we pose a two-step approach to carry out the evaluation. The first step of this approach is to establish an estimation of the probabilistic ground truth through a function that uses the labels of the crowd. Then, once this is established, in the second step an ordinary estimator (one that requires the availability of the ground truth) of the evaluation measure able to deal with the established estimation of the probabilistic ground truth is used, since now the values associated to each instance are no longer bounded to the set $\{0, 1\}$, as they are within the range $[0, 1]$.

The establishment of the estimation of the probabilistic ground truth of the first step can be accomplished, for instance, through the maximum likelihood estimation of the probability distribution. Alternatively, a Bayesian estimation can be used to introduce a priori knowledge.

3.3.1. Probabilistic Ground Truth (PGT) method

In this case, for each instance, we propose to use the labels to establish an estimation of the probabilistic ground truth through the maximum likelihood estimation with the Jeffreys-Perks correction [52, 53] of the probability distribution. The Jeffreys-Perks correction has been selected so as to take into account the relative amount of information issued for each instance (in order to induce more uncertainty in the instances with fewer labels). In this context, it simplifies to (where C is the class variable, i specifies the current instance, $\boldsymbol{x}_i$ is the vector of predictive variable values for the current instance, n represents the amount of annotators, j is the current annotator and l_{ij} is the label generated by the j-th annotator for the i-th instance):

$$\hat{P}(C=1|\boldsymbol{x}_i) = \frac{\sum\limits_{j=1}^{n} l_{ij} + \frac{1}{2}}{n+1}.$$

It seems advisable to remember that the AUC can be seen as a way of measuring the distance between a specific ordering of binary elements and the perfect ordering of the same binary elements [54]. But the AUC of a given classifier regarding the estimation of the probabilistic ground truth cannot be measured directly since now an ordering of values in the domain $[0, 1]$ has to be dealt with. Consequently, it seems convenient to check if there exist generalizations or extensions of the AUC that can tackle that kind of situation. One option is the extension consisting of the volume under the ROC surface for multi-class problems [55]. Another option consists of the generalization of the AUC for multiple class classification problems through the averaging of pairwise comparisons [54]. These two options are conceived for problems in which there is no ordinal relationship between classes, but in the problem at hand, the ordinal relationship exists and matters. Fortunately, an assessment that takes into account the ordinal relationship of the values in the domain $[0, 1]$ can be made. Specifically, it can be made thanks to the metric selected in this section, which is based on the normalized Kendall-Tau distance for permutations [34] (specific orderings of the elements of a given set), generalized for multi-permutations [31, 32] (permutations of the elements from a multi-set, which is a set with repeated elements), which in fact, supposes a generalization of the estimation of the AUC. Moreover, when the multi-permutations have only two different values, the AUC is equal to one minus the normalized Kendall-Tau distance between that multi-permutation and the multi-permutation of the same elements that lets the elements be ordered decreasingly (which is the value computed by the metric). This connection between them is already known [33]. Finally, in order to better illustrate how this method works, its algorithm is shown in Figure 3 , in which l_i represents the estimated probabilistic ground truth for instance i.

```
Algorithm 3: PGT
    Input: $\boldsymbol{q}$ and $L$
    Output: measure
    begin
        for $i \leftarrow 1$ to $m$ do
            $l_i \leftarrow$ estim_prob($L_{i\cdot}$)
        end
        measure $\leftarrow$ 1 - norm_kendall_tau_dist($\boldsymbol{q}, \boldsymbol{l}$)
        return measure
    end
```

Figure 3: Algorithm of the PGT method.

4. Experimentation

In the experimentation, our aim is to profile the capability of the three proposed estimators to perform a sensible selection of models in different configurations of problems in which crowdsourced data are used. To do so, we have randomly generated a ground truth and then we have simulated the crowdsourced labels and the outputs of a set of models. In addition, the supervised case of the instances is simulated to use its performance as a reference. Next, the three evaluation methods proposed, which use the labels of the annotators to perform the evaluation, and the true evaluation, which uses the ground truth to perform the evaluation, are computed. Finally, the disagreements of the three evaluation methods and of the supervised case regarding the ground truth are measured. It is worth mentioning that we have used only simulated problems in order to compare the three methods with the evaluation in the presence of classified instances.

Briefly, the process is as follows, the probabilistic ground truth of the m instances, $\boldsymbol{p} = (p_1, \cdots, p_m)$, where p_i represents $P(C = 1|\boldsymbol{x}_i)$, is randomly generated and then, based on this, the labels of the annotators (represented with the matrix L, of m rows each one related to a different instance and of n columns each one related to a different labeler) and the outputs of the classifiers (represented with the matrix Q, of m rows each one related to a different instance and of k columns each one related to a different classifier) are simulated.

Next, the true evaluation of the classifiers is computed using $\boldsymbol{p}$ and Q, while the three proposed evaluation methods are computed using Q and L. The supervised case that serves as a reference is achieved through a sample of the true distribution $\boldsymbol{p}$, which allows the labeling of the instances. Having the labels of the instances and the outputs of the classifiers, the standard AUC is estimated. Finally, the evaluation of each evaluation method (and, similarly, of the supervised case) is made in terms of the similarity between their rankings of classifiers and the ranking of classifiers according to the true evaluation, in consonance with what was indicated in Section 2. The general process we have designed for the experimentation is summarized in Figure 4. It is worth noting that, although we use that general framework for binary classification, the process summarized in Figure 4 is not bounded to binary classification and can be applied to multiclass classification problems, if the necessary changes to enable the calculations carried out inside each box of Figure 4 for multiclass classification problems are made. In the simulation conducted in this work, different configurations specified through parameters are tested, running each of them 100 times. Namely, the general process shown in Figure 4 is run 100 times per configuration.

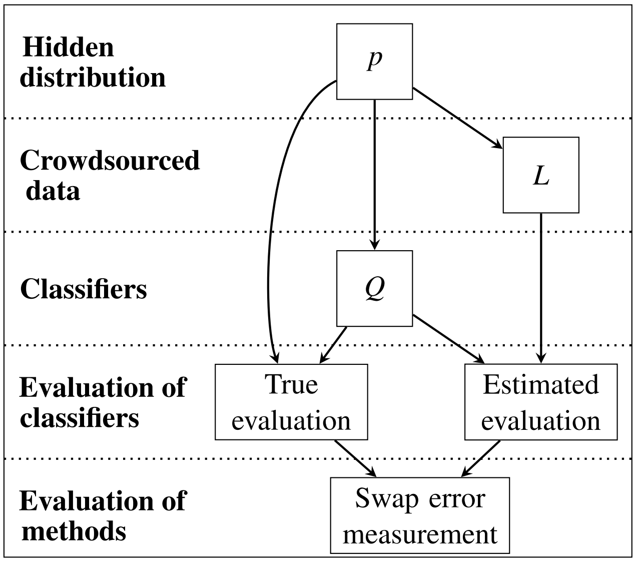

Figure 4: General process of the experimentation.

Until now, the general process of the experimentation has been exposed so as to give a global view of it. Henceforth, we proceed to sequentially expose how each of the 6 steps that appear in Figure 4 is computed.

1. $\boldsymbol{p}$: Values for $m = 1000$ instances are generated per run with a non-informative approach. This non-informative approach is achieved by generating the values $p_i = P(C = 1|\boldsymbol{x}_i)$ for the different instances through a Beta distribution with parameter values $\alpha = 1$ and $\beta = 1$.
2. L: This step is the one in which the crowdsourced data joins the experimentation through the labels issued by the annotators. In order to test different configurations, we have developed a set of parameters that enable the control of the ratio of annotators to instances, the distribution of the qualities of the annotators and the degree of sparseness of the matrix L. Those parameters are the *expected amount of labels per annotator* $\bar{l}_j$, the *expected amount of labels per instance* $\bar{l}_i$, the average true positive rate of the annotators μ_{TPR} and the average false positive rate of the annotators μ_{FPR}.

   The parameters $\bar{l}_j$ and $\bar{l}_i$, given a fixed m, allow the control of the ratio of annotators to instances and the degree of sparseness of the matrix L. This control can be expressed through the equivalences $\bar{l}_j = m \cdot t$ and $\bar{l}_i = n \cdot t$, with t being the probability of a random annotator labeling a random instance, and n being the amount of annotators. In the experimentation, the parameters $\bar{l}_i$ and $\bar{l}_j$ take the values 3, 5, 9, 15, 27, and 45.

The parameters μ_{TPR} and μ_{FPR} allow the control of the distribution of the qualities of the annotators, which are expressed in terms of sensitivity or true positive rate (TPR, which is the proportion of positive instances that are labeled correctly) and in terms of one minus the specificity or false positive rate (FPR, which is the proportion of negative instances that are labeled incorrectly). Specifically, the TPR and the FPR of each annotator are sampled from the distributions Beta$(2, \beta) \cdot 0.5 + 0.5$ and Beta$(\alpha, 2) \cdot 0.5$ respectively, being β and α:

$$\beta = \frac{4 \cdot \mu_{\text{TPR}} - 4}{1 - 2 \cdot \mu_{\text{TPR}}}, \qquad \alpha = \frac{4 \cdot \mu_{\text{FPR}}}{1 - 2 \cdot \mu_{\text{FPR}}}.$$

Consequently, in the experimentation, the TPR and the FPR of each annotator are defined in the bounded intervals $[0.5, 1]$ and $[0, 0.5]$ respectively. It is worth noting that, for a given annotator, the further they are from 0.5, the better the quality of the annotator. In Figure 5, the plane in which the qualities of the annotators are represented in terms of their TPR and FPR is shown. In order to give a brief notion of how skilled the annotators are depending on which region of the plane shown in Figure 5 they are in, in that figure we have divided the domain of definition of the qualities of the annotators in this experimentation into 9 equally sized regions. Each of those 9 regions has been labeled with a representative description (‘very bad”, “bad”, “skewed”, “average”, “good” or “very good’) of the qualities of the annotators belonging to the given region. For instance, the best annotators will tend to have TPR values close to 1 and FPR values close to 0, while the worst of this experimentation will tend to have TPR and FPR values close to 0.5. Besides, skewed annotators will tend to have either TPR values close to 1 and FPR values close to 0.5, thus issuing a lot of ones, or TPR values close to 0.5 and FPR values close to 0, thus issuing a lot of zeros.

Specifically, in our experimentation, the parameters μ_{TPR} and μ_{FPR} that control the qualities of the annotators, jointly take the pairs of values $(0.525, 0.475)$, $(0.6, 0.4)$, $(0.7, 0.3)$, $(0.8, 0.2)$, and $(0.9, 0.1)$, drawing the cases to which we refer to as “Extreme”, “Bad”, “Average”, “Good” and “Outstanding” respectively. The theoretical distributions of the qualities of the annotators in terms of TPR and FPR for those five cases can be calculated. In order to do so, first the α and β parameters of the distributions of the TPR and the FPR of the annotators are computed, considering that, for each of the five cases, different μ_{TPR} and μ_{FPR} values are used. Once the values of those parameters are computed, the probability density functions of the TPR and the FPR become available. Finally, for each case, the joint probability density function of the TPR and the FPR, which describe the distribution of the qualities of the annotators, is computed as the multiplication between the probability density functions of the TPR and the FPR of that case. In order to give some hints regarding the shapes of the theoretical distribution of each case, the Figures 6, 7, 8, 9 and 10 are displayed. In them, in each region, the probability to sample a quality for an annotator whose TPR and FPR values can be represented as a point within that region is numerically represented.

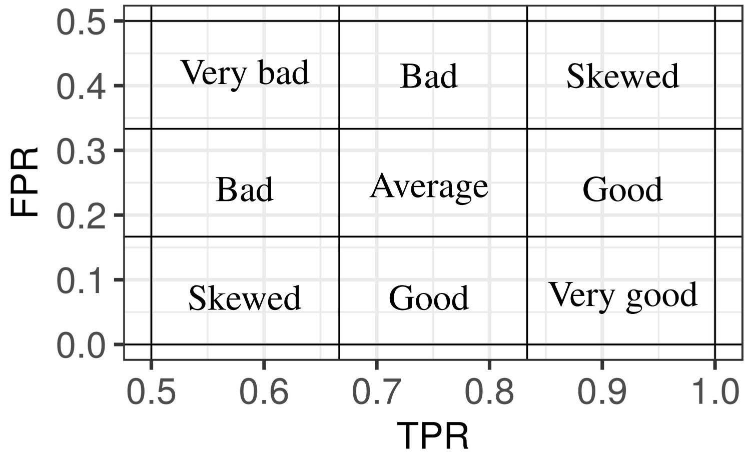

Figure 5: Names given to the annotators by their qualities.

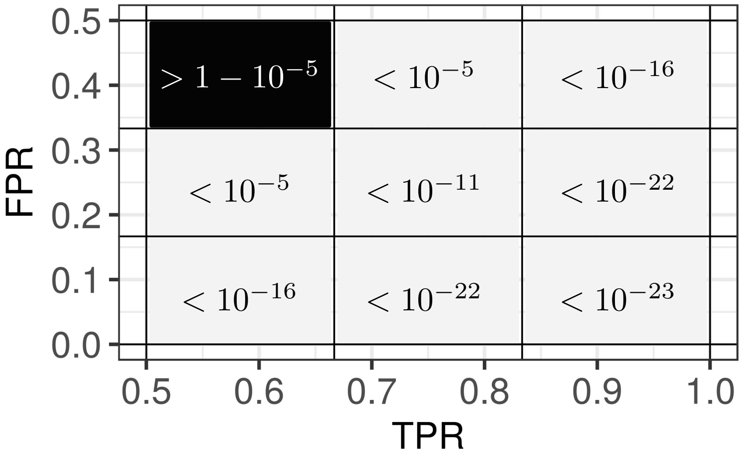

Figure 6: Theoretical distributions of the qualities of the annotators when $\mu_{\text{TPR}} = 0.525$ and $\mu_{\text{FPR}} = 0.475$ ("extreme" case).

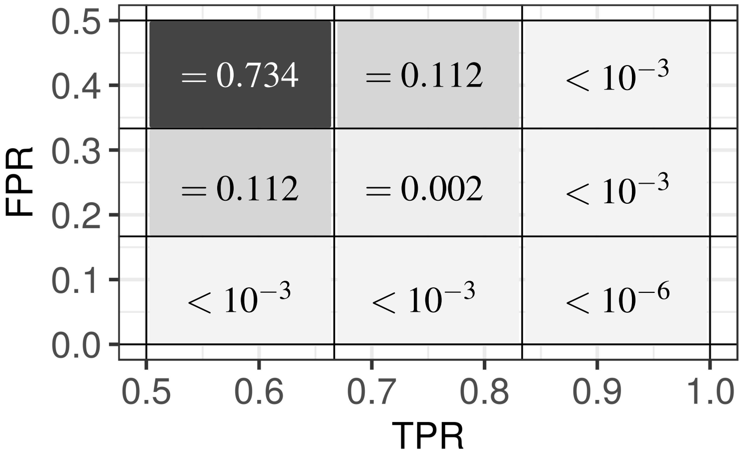

Figure 7: Theoretical distributions of the qualities of the annotators when $\mu_{\text{TPR}} = 0.6$ and $\mu_{\text{FPR}} = 0.4$ ("bad" case).

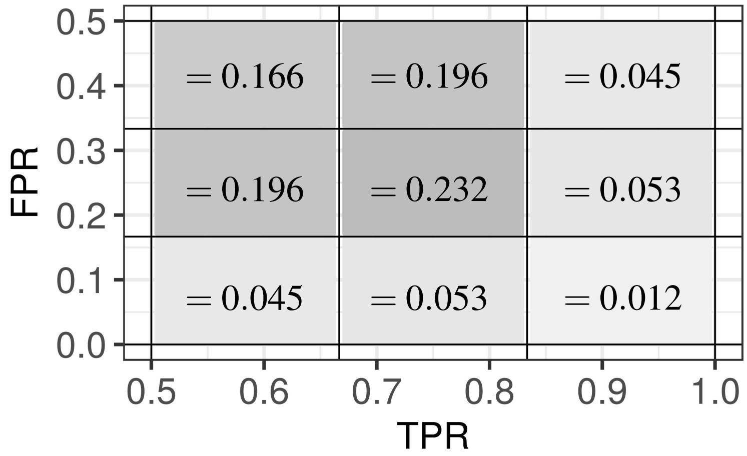

Figure 8: Theoretical distributions of the qualities of the annotators when $\mu_{\text{TPR}} = 0.7$ and $\mu_{\text{FPR}} = 0.3$ ("average" case).

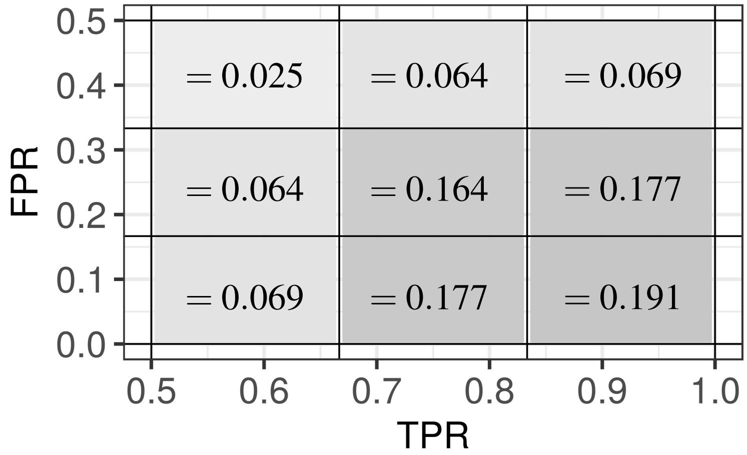

Figure 9: Theoretical distributions of the qualities of the annotators when $\mu_{\text{TPR}} = 0.8$ and $\mu_{\text{FPR}} = 0.2$ ("good" case).

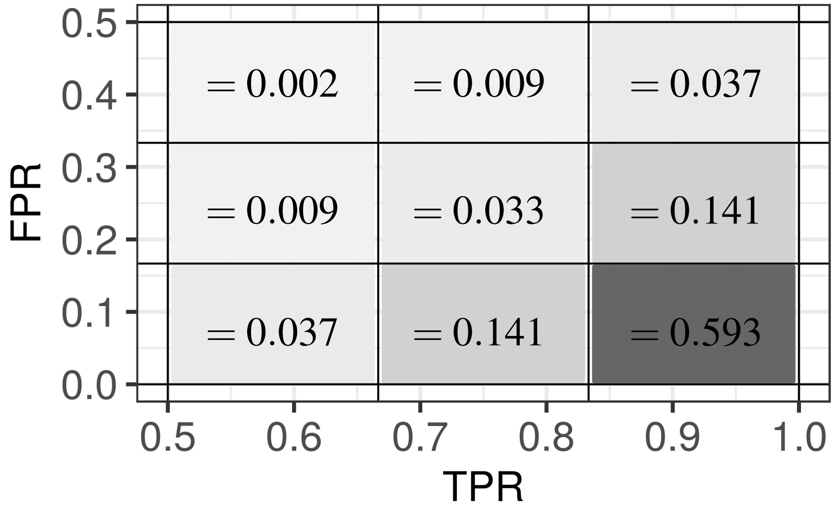

Figure 10: Theoretical distributions of the qualities of the annotators when $\mu_{\text{TPR}} = 0.9$ and $\mu_{\text{FPR}} = 0.1$ ("outstanding" case).

Finally, the generation of L is completed the following way. For a given j-th annotator, once the TPR and the FPR associated to the annotator are sampled (denoted as TPR_j and FPR_j, respectively) the label value for the i-th instance, l_{ij}, can be randomly sampled given the distribution:

$$P(l_{ij} = 1) \quad = \quad p_i \cdot \text{TPR}_j + (1 - p_i) \cdot \text{FPR}_j.$$

3. Q: Specifically, 100 different classifiers are simulated, therefore $k = 100$. Note that these 100 classifiers may also represent the same kind of classifier, taking 100 different parameter settings. Each of those predictions is considered to be a disturbed version of the (probabilistic) ground truth since the classifiers are supposed to learn to predict the ground truth. In particular, letting a given value of d denote a specific classifier and its associated disturbance, with $d \in \{1, \ldots, 100\}$, the output of the d-th classifier for the i-th instance is specified as:

$$
\begin{aligned}
q_{id} \quad &= \quad p_i \cdot \mathcal{U}(1 - 0.45 \cdot \frac{d-1}{99}, 1) + \\
&\qquad (1-p_i) \cdot \mathcal{U}(0, 0.45 \cdot \frac{d-1}{99}).
\end{aligned}
$$

The disturbance introduced in the outcome of the d-th classifier when it is dealing with the i-th instance can be quantified in terms of the expected disagreement between p_i and q_{id}. Namely:

$$
\mathbb{E}(|p_i - q_{id}||p_i, d) \quad = \quad |(2 \cdot p_i - 1) \cdot \frac{0.45 \cdot (d-1)}{2 \cdot 99}|.
$$

As can be seen, for the classifiers 1 to 100, on average, the higher the value of d, the stronger the expected disturbance of the outcomes of the associated classifier (except for $p_i = 0.5$). Besides, as the disturbance increases, the similarity between the ordering of the instances according to $\boldsymbol{p}$ and the ordering of the instances according to $Q_{\cdot d}$ ($Q_{\cdot d}$ being the outcomes of the d-th classifier) tends to decrease, similarity which will be used as the assessment of the goodness of the classifiers (see true evaluation step). Consequently, as d increases the goodness of its associated classifier tends to decrease.

It is worth mentioning that the reason why 100 classifiers of different degrees of goodness are used is to make the problem of ranking them correctly a difficult one. In fact, the problem has been made a difficult one in order to ease the observance of differences between the performances of the evaluation methods. Namely, if the problem had been made too easy, the three evaluation methods would have probably performed well and would have probably shown small differences between them. As an illustration of the difficulty (in addition to the computation of the supervised case), the mean and the standard deviation through the different runs of the maximum relative difference of the true evaluation (see true evaluation step for details of its computation) between adjacent classifiers (those where there is a difference between their d values is one) are only 1.4% and 0.3% respectively. Another possible illustration of the difficulty consists of the boxplots corresponding to the true evaluations of the classifiers, boxplots in which can be clearly seen that ranking them correctly is a difficult task. For the sake of simplicity, we leave such boxplots in the supplementary material.

4. The true evaluation: Using Q and $\boldsymbol{p}$, the true evaluation is computed for every classifier. Specifically, the true evaluation for the d-th classifier is based on the similarity between the rankings of instances derivable from $\boldsymbol{p}$ and $Q_{\cdot d}$. In order to assess that similarity the aforementioned generalization of the AUC is used. Namely, we compute the expression consisting of one minus the normalized Kendall-Tau distance between two permutations (or multi-permutations) [31, 32, 34] of the elements of $\boldsymbol{p}$, the first one being ordered by $\boldsymbol{p}$ and the second one being ordered by $Q_{\cdot d}$. Briefly, what is computed is one minus the proportion of pairs of instances that appear in different order among them (which one is ranked before the other) when the ordering of the instances is according to $\boldsymbol{p}$ and when the ordering of the instances is according to $Q_{\cdot d}$. The selection of such a measure to assess the similarity between pairs of rankings is motivated by its interesting properties. On the one hand, it considers that each possible pair of elements (instances in this case) of the rankings has the same relevance when assessing its consistency or inconsistency in terms of how are they ordered in the two different rankings, it being aseptic in that matter. On the other hand, it is shares the advantages of the AUC that were exposed in Section 2.

   As a matter of a fact, since the hidden distribution is known, more information is available than when the "true" labels are known. In other words, there is no need to fix a threshold value to generate label values for $\boldsymbol{p}$. Finally, once the true evaluation of every classifier is available, a ranking of them according to their true evaluations is elaborated.

5. The estimated evaluation: Using Q and L, the three estimated evaluations are computed for each classifier through the use of the three methods exposed in this paper, DGT, SGT and PGT. This allows us to elaborate three rankings of the classifiers, each of them being concordant with the performances that the classifiers obtain according to a different method of the three exposed ones.

6. The swap error measurement: As exposed in section 2, our aim is to perform a proper model selection, which basically consists of identifying correctly, for

every pair of classifiers, which one of the pair has achieved a better or worse performance than the other. That is essentially to build a ranking of models that ranks every one of them correctly. Consequently, we focus on assessing the three methods proposed in this work in terms of how similar their three associated rankings of classifiers are to the ranking of classifiers associated to the true evaluation. For the assessment of how similar a pair of rankings are, this time, instead of assessing it in terms of one minus the normalized Kendall-Tau distance for permutations (or multi-permutations), we choose to express it directly in the normalized Kendall-Tau distance for permutations (or multi-permutations). Namely, instead of expressing how similar two rankings are by assessing their similarity (and therefore the goodness of the evaluation method), this time we choose to express how similar two rankings are by assessing their dissimilarity (and therefore the error of the evaluation method). We refer to such error as *swap error*, for which, by agreement, we consider rankings with swap error < 0.1, $0.1 \leq$ swap error < 0.25 and $0.25 \leq$ swap error as high, middle and low quality solutions, respectively, given that the best possible swap error is 0, the worst possible swap error is 1 and the expected swap error between two rankings generated at random is 0.5.

5. Results

In this section, we present the results obtained for the three methods to perform the evaluation presented in Section 3 (DGT, SGT and PGT). To compare the qualities of the three methods in terms of the swap errors they made regarding the true evaluation in the different configurations, a boxplot and a scatter plot is generated per configuration. In each boxplot there is one box per evaluation method that shows the swap error that each evaluation method makes regarding the true evaluation. In addition, each boxplot incorporates the horizontal line corresponding to the average performance of the supervised approach (which has the quartile values $Q_1 = 0.0806$, $Q_2 = 0.0873$ and $Q_3 = 0.0930$) as a way to assess the performance of the evaluation methods. The content shown and the distribution of the configurations in the scatter plots its analogous, replacing only the box and whiskers by the corresponding dots. For the sake of simplicity, in this paper we present a representative subset of the boxplots, in which the same general trends can be appreciated, thus leaving the full set of boxplots and scatter plots as supplementary material. The selected subset of boxplots is shown in Figures 11, 12, 13, 14 and 15. In each of these figures, a grid of 9 boxplots corresponding to configurations that take the same μ_{TPR} and the same μ_{FPR} values are displayed. Each of the columns in the grids of plots implies a different value of $\bar{l}_j$, while each of the rows implies a different value of $\bar{l}_i$. Besides, as $\bar{l}_j$ increases, the ratio of annotators to instances decreases, while as $\bar{l}_i$ increases, the ratio increases.

To start with, it can be seen in Figures 11, 12, 13, 14 and 15 that for almost each evaluation method in each configuration the average swap error achieved is less than the expected swap error of an evaluation method which ranks the classifiers at random (0.5). Moreover, in a large majority of the configurations the average swap error is far better than the expected swap error of a random evaluation method. In addition, the very few configurations in which the obtained results are comparable to a random behavior are those in which it both happens that the labelers are very unskilled and that there are very few labels per instance. Consequently, the empirical results support the usefulness of the proposed methods.

In the figures of matrices of boxplots, all the evaluation methods improve (reduce) their swap error as $\bar{l}_i$ increases, decreasing both the mean and the variance of the swap error values. This makes sense, since there is more information available per instance as the $\bar{l}_i$ axis is traveled.

It can be seen that varying only the parameter $\bar{l}_j$ only affects the SGT method effectively, improving its swap error as $\bar{l}_j$ increases. However, this effect is weaker than the one caused by the variation of $\bar{l}_i$. Nonetheless, the effect is strong enough to make the SGT vary from being similar to the worst method when there are few labels per annotator to being similar to the best one when there are many labels per annotator. The source of this effect is that, as $\bar{l}_j$ increases, there are on average more labels issued by each individual annotator, i.e., more information, more points, available to perform the AUC estimations. So, the SGT estimation improves as the estimated AUC values for the annotators improve.

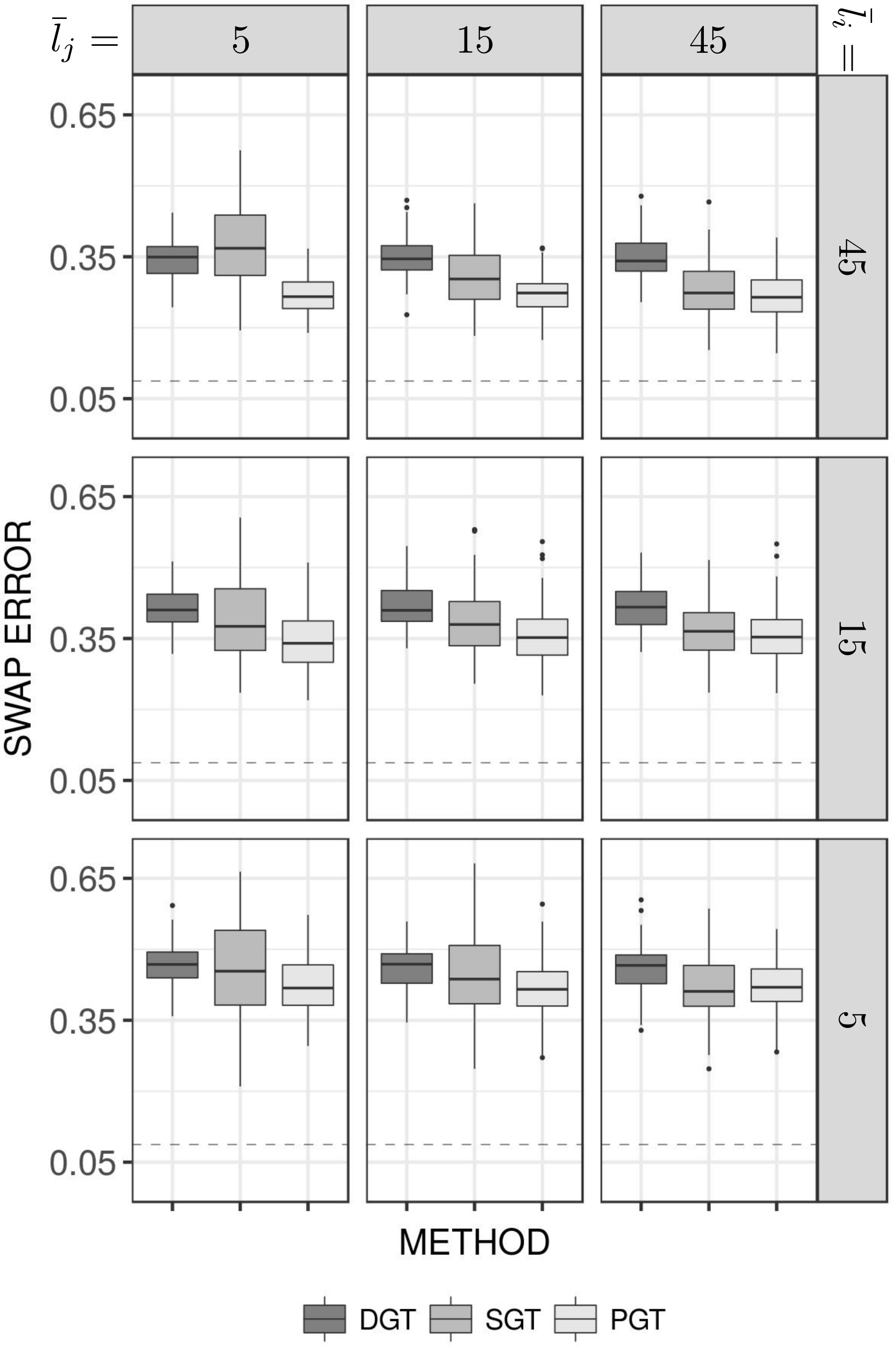

Figure 11: Boxplots of the swap errors the different methods achieve regarding the true evaluation in terms of rankings of classifiers, for the "extreme" case in which $\bar{l}_j$ and $\bar{l}_i$ are each either 5, 15 or 45.

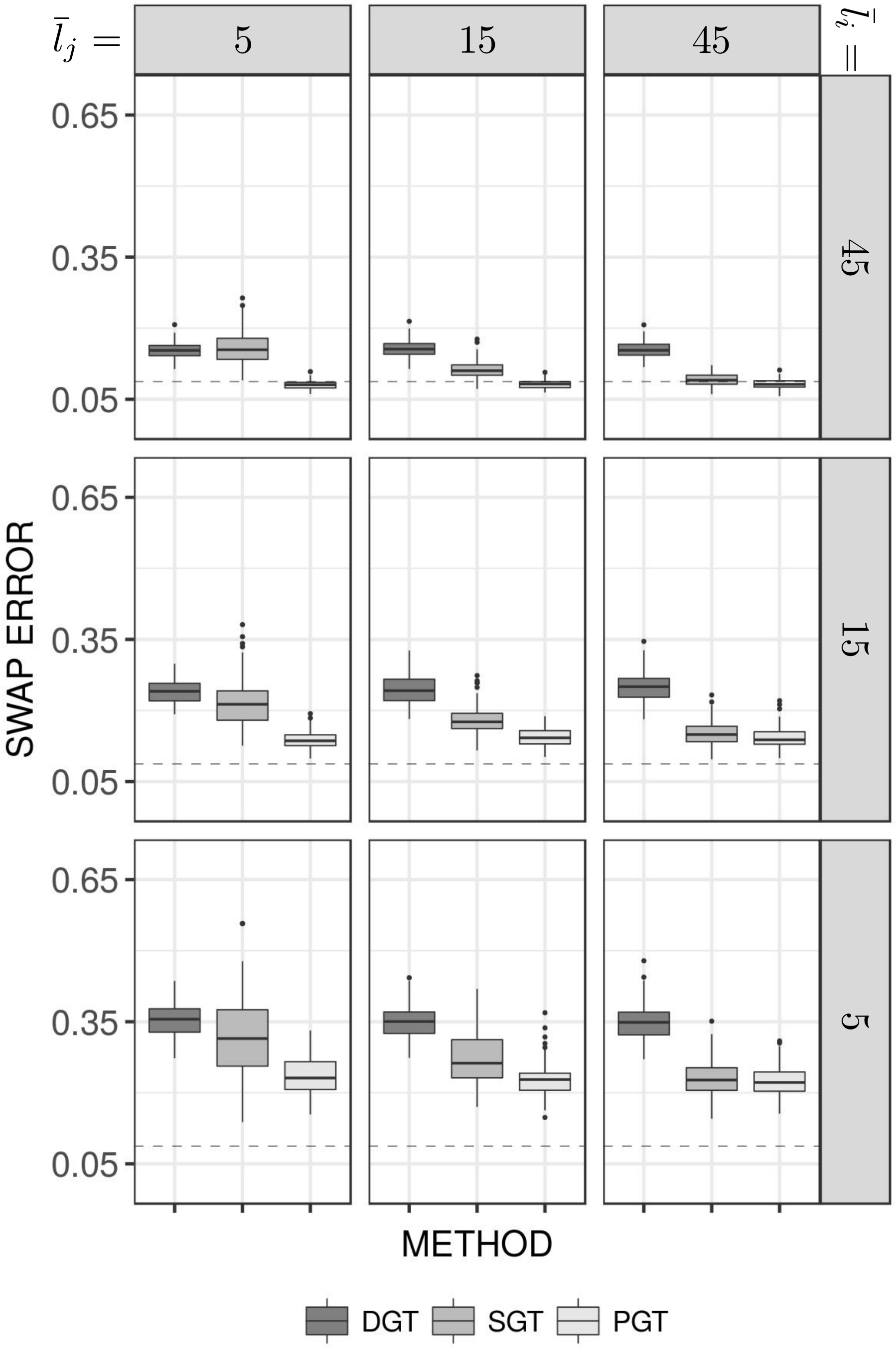

Figure 12: Boxplots of the swap errors the different methods achieve regarding the true evaluation in terms of rankings of classifiers, for the "bad" case in which $\bar{l}_j$ and $\bar{l}_i$ are each either 5, 15 or 45.

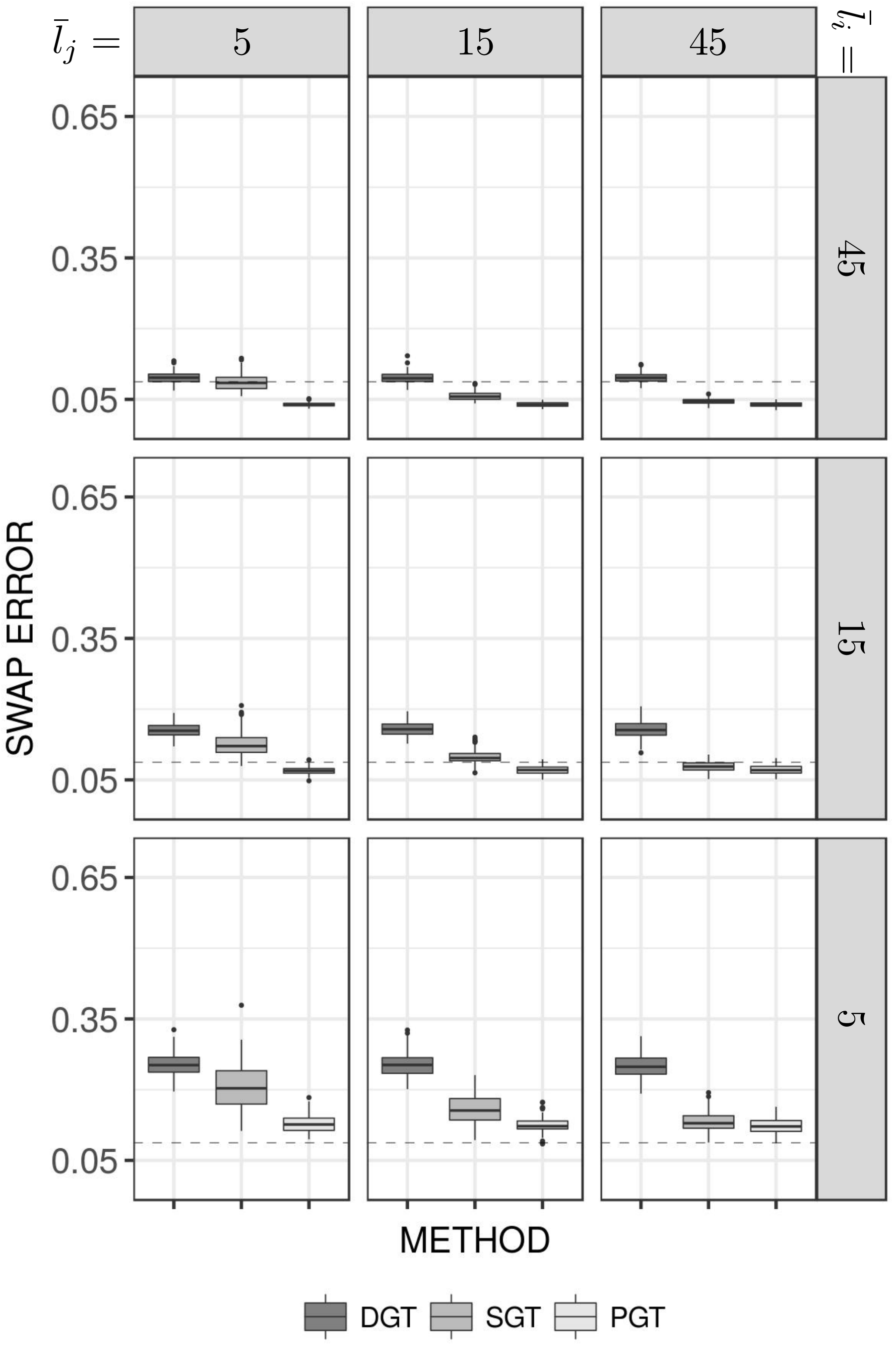

Figure 13: Boxplots of the swap errors the different methods achieve regarding the true evaluation in terms of rankings of classifiers, for the "average" case in which $\bar{l}_j$ and $\bar{l}_i$ are each either 5, 15 or 45.

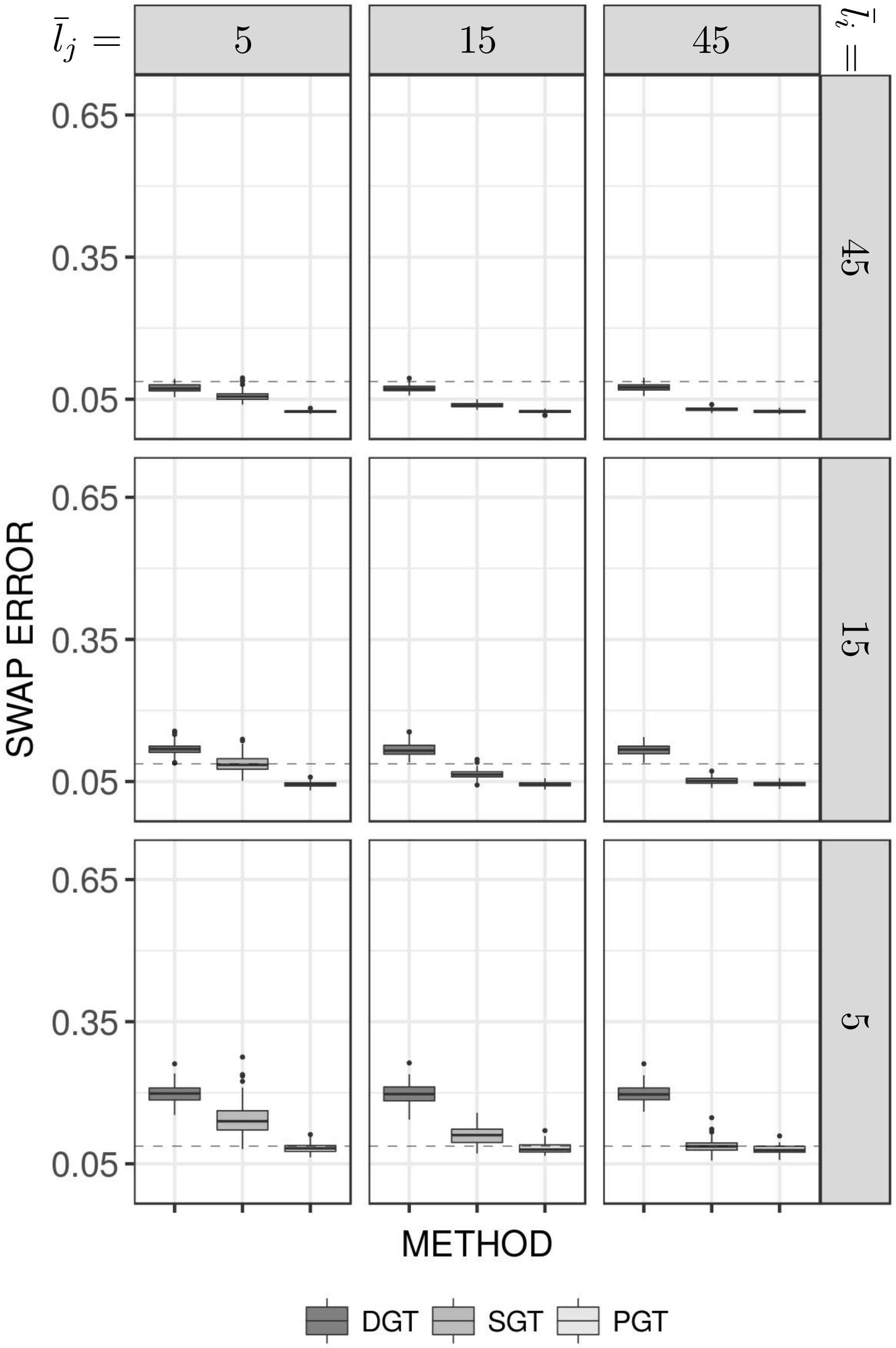

Figure 14: Boxplots of the swap errors the different methods achieve regarding the true evaluation in terms of rankings of classifiers, for the "good" case in which $\bar{l}_j$ and $\bar{l}_i$ are each either 5, 15 or 45.

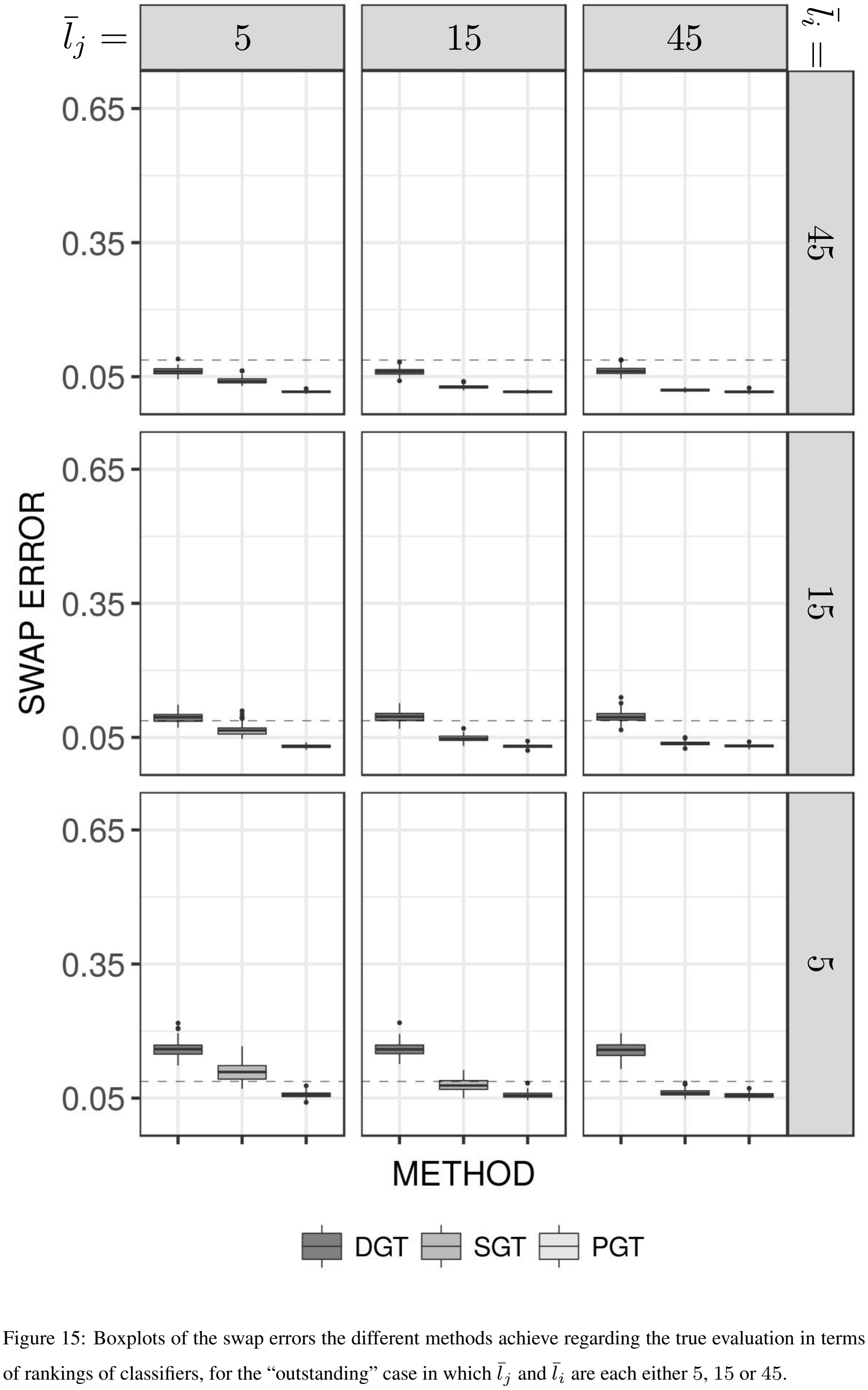

Figure 15: Boxplots of the swap errors the different methods achieve regarding the true evaluation in terms of rankings of classifiers, for the "outstanding" case in which $\bar{l}_j$ and $\bar{l}_i$ are each either 5, 15 or 45.

PGT offers, on average, the best results, achieving in general the lowest mean swap error. Additionally, PGT behaves especially well in comparison to the other two methods in the synthetic problems that represent better crowdsourced data problems, in which it is likely that each annotator labels only a few instances [9, 23, 27]. Contrarily, DGT offers, on average, the worst results, being unable to match PGT in any configuration. Finally, SGT can issue, depending on the specific configuration, results that match PGT, results that match DGT or results that lie in the interstice between PGT and DGT.

In terms of variability, it must be said that SGT stands out for being the most unstable one overall. This difference in variability is especially high in configurations where there are 15 labels or fewer per annotator.

Finally, the cases are described in detail (for a fully in-depth detailed view see supplementary material) regarding the qualities of the annotators in terms of the average swap error:

- "Extreme" case: All methods achieve low quality solutions for every configuration.

- "Bad" case: DGT ranges between low and middle quality solutions, while SGT and PGT range between low and high quality solutions. With 45 labels per instance, in mean terms, PGT beats the average performance of the approximation through a supervised approach.

- "Average" case: DGT and SGT range between low and high quality solutions, while PGT ranges between middle and high quality solutions. With 15 or more labels per instance, in mean terms, PGT always beats the supervised approach and SGT beats it in 61.11% of those cases.

- "Good" case: All the methods range between middle and high quality solutions. With five labels or more per instance, on average terms, PGT always beats the supervised approach, while SGT does it in 70% of the configurations. With 45 labels or more, all the methods beat the supervised approach in mean terms.

- “Outstanding” case: Again, all the methods range between middle and high quality solutions. This time PGT, in mean terms, always beats the supervised approach, while SGT requires 15 labels or more to beat it always, and DGT requires 27 or more to do the same.

6. Conclusions

In this paper we have tackled the problem of the evaluation and model selection in binary classification problems on crowdsourced data contexts in which the ground truth is not available, which, as far as we know, we are the first to deal with it explicitly. To start with, the problem at hand is defined and formalized. Secondly, three general approaches to undertake the task have been identified, allowing each one to outline the ground truth according to a specific conception of its nature. In addition, a particular evaluation method belonging to each approach and which is capable of ranking classifiers in crowdsourced data contexts without the ground truth, through an estimation of their AUCs, is specified and described in detail. Next, the three proposed estimators are tested in an extensive synthetic experimentation, which is composed of many configurations in order to favor the profiling of their global performance in plausible scenarios of the model selection problem in crowdsourced data contexts. In the experimentation, in order to achieve our main objective, which is the assessment and comparison of evaluation methods capable of performing the model selection in such conditions (together with the proposal of the methods themselves), the framework summarized in Figure 4 has been used. Specifically, in the experimentation, we have proposed the use of normalized Kendall-Tau distance as a sensible measure with which to quantify how well or badly the proposed estimators perform in order to enable their comparison.

Regarding the designed evaluation methods, the DGT method can be seen as a discretized version of the PGT method. That discretization implies a loss of detail strong enough to increase its swap error, a loss that hinders discerning classifiers with similar true evaluation values.

The SGT method highly depends on how many labels the annotators issue on average. Consequently, it seems more suited for problems in which, normally, the annotators are able to issue many labels, which is not usually the case of crowdsourced labels. Rather, SGT seems more suited for problems in which a committee of experts labels data, because in such problems the annotators tend to issue more labels. Besides, although in the results a region of the space of problems in which SGT behaves consistently better than PGT cannot be seen, the results hint that such a region may exist, given the tendency SGT shows when the labels per annotator tends to increase.

According to the results, PGT is the best evaluation method among the three proposed, because, in general, in comparison with the other methods it reaches the lowest mean swap error. In addition to the results it achieved, the PGT method presents a set of advantages. To start with, PGT is insensitive (together with DGT) to the amount of labels issued per annotator. Besides, the process of the estimation of the probabilistic ground truth allows great flexibility. On the one hand, the reliability of the labelers can be assessed using weights or Bayesian a prioris. On the other hand, in order to smooth the influence of each instance regarding the amount of information issued for each one, i.e., the amount of labels per instance, corrections can be applied, such as, for instance, the Laplace correction or the Jeffreys-Perks correction. Lastly, although it is not a situation tested here, compared with SGT, PGT has the advantage of being applicable in scenarios where which annotator each label belongs to is unknown.

7. Recommendations and further studies

In the previous section, we have exposed the conclusions drawn from the work done. Taking these conclusions into account, our main recommendation is to use the PGT method for model selection in problems with crowdsourced data, given the better results it achieves and the several advantages it has. However, it should be taken into account that the reliability of its outcomes depends directly on the distribution of the quality of the annotators (as do the outcomes of DGT and SGT). Consequently, its outcomes should be considered reliable in proportion to the expected reliability of the annotators.

Regarding further studies, we now identify several ideas that seem interesting to carry out in the near future. To start with, one interesting task to be performed is to extend the experimentation to explore different regions of the space of problems. Specifically, extending the experimentation to test configurations that match scenarios of committees of experts seems very interesting. The main reason is that the current results hint that, in such problems, SGT may behave better than PGT, although it remains to be researched. Another interesting idea consists of not only comparing the different methods exposed here in synthetic problems, but also in real problems (using the general process represented in Figure 4), seeking to consolidate the knowledge acquired regarding the behaviors of DGT, SGT and PGT. To tackle that task, those real problems should have different predictive variables, an available ground truth, an available matrix of labels issued by annotators and an available matrix of classifications made by trained classifiers. Another area worth exploring consists of the development of alternative evaluation methods, seeking to improve the results of the three methods exposed in this paper, so as to achieve better estimations of the true evaluations of the classifiers. Finally, another idea to carry out in the future consists of extending the work done here to non-binary classification problems, so as to provide evaluation methods capable of performing the selection of models in such a context.

Declarations of interest

None.

Acknowledgements

This work is partially supported by the Basque Government (IT609-13 and Elkartek BID3A) and the Spanish Ministry of Economy and Competitiveness (TIN2016-78365-R) and a University-Society Project 15/19 (Basque Government and University of the Basque Country UPV/EHU). Ari Urkullu has been supported by the Basque Government through a predoctoral grant (PRE_2013_1_1313, PRE_2014_2_87, PRE_2015_2_0280 and PRE_2016_2_0314). Aritz Pérez is supported by the Basque Government through the BERC 2018-2021 and the ELKARTEK programs, and by the Spanish Ministry of Economy and Competitiveness MINECO through BCAM Severo Ochoa excellence accreditation SEV-2017-2018. Aritz Pérez is also supported through TIN2017-82626-R funded by (AEI/FEDER, UE).

References

[1] S. Albarqouni, C. Baur, F. Achilles, V. Belagiannis, S. Demirci, N. Navab, Aggnet: deep learning from crowds for mitosis detection in breast cancer histology images, IEEE transactions on medical imaging 35 (5) (2016) 1313–1321.

[2] C. Arteta, V. Lempitsky, A. Zisserman, Counting in the wild, in: European conference on computer vision, Springer, 2016, pp. 483–498.

[3] K. Benoit, D. Conway, B. E. Lauderdale, M. Laver, S. Mikhaylov, Crowd-sourced text analysis: Reproducible and agile production of political data, American Political Science Review 110 (2) (2016) 278–295.

[4] C. Callison-Burch, Fast, Cheap, and Creative: Evaluating Translation Quality Using Amazon's Mechanical Turk, in: Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing, Association for Computational Linguistics, Singapore, Singapore, 2009, pp. 286–295.

[5] X. Chen, Q. Lin, D. Zhou, Statistical Decision Making for Optimal Budget Allocation in Crowd Labeling, Journal of Machine Learning Research 16 (2015) 1–46.

[6] J. Hernández-González, I. Inza, J. A. Lozano, Learning from Crowds in Multidimensional Classification Domains, in: Conference of the Spanish Association for Artificial Intelligence, Springer, Madrid, Spain, 2013, pp. 352–362.

[7] J. Hernández-González, I. Inza, J. A. Lozano, Multidimensional Learning from Crowds: Usefulness and Application of Expertise Detection, International Journal of Intelligent Systems 30 (3) (2015) 326–354. doi:10.1002/int.21702.
URL https://doi.org/10.1002/int.21702

[8] H. Hu, Y. Zheng, Z. Bao, G. Li, J. Feng, R. Cheng, Crowdsourced poi labelling: Location-aware result inference and task assignment, in: Data Engineering (ICDE), 2016 IEEE 32nd International Conference on, IEEE, 2016, pp. 61–72.

[9] V. C. Raykar, S. Yu, L. H. Zhao, G. H. Valadez, C. Florin, L. Bogoni, L. Moy, Learning from crowds, Journal of Machine Learning Research 11 (2010) 1297–1322.

[10] P. Smyth, U. M. Fayyad, M. C. Burl, P. Perona, P. Baldi, Inferring Ground Truth from Subjective Labelling of Venus Images, in: G. Tesauro, D. Touretzky, T. Leen (Eds.), Advances in Neural Information Processing Systems 7, MIT Press, Denver, USA, 1995, pp. 1085–1092.

[11] G. Van Horn, S. Branson, S. Loarie, S. Belongie, P. Perona, G. Van Horn, O. Mac Aodha, Y. Song, Y. Cui, C. Sun, et al., Lean multiclass crowdsourcing, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2018, pp. 2714–2723.

[12] Y. Yan, G. M. Fung, R. Rosales, J. G. Dy, Active learning from crowds, in: Proceedings of the 28th International Conference on Machine Learning, ICML11, International Machine Learning Society (IMLS), Bellevue, USA, 2011, pp. 1161–1168.

[13] J. Zhang, X. Wu, V. S. Sheng, Imbalanced Multiple Noisy Labeling, IEEE Transactions on Knowledge and Data Engineering 27 (2) (2015) 489–503. doi:10.1109/TKDE.2014.2327039.
URL https://doi.org/10.1109/TKDE.2014.2327039

[14] Y. Zhang, X. Chen, D. Zhou, M. I. Jordan, Spectral methods meet em: A provably optimal algorithm for crowdsourcing, The Journal of Machine Learning Research 17 (1) (2016) 3537–3580.

[15] J. Zhang, V. S. Sheng, T. Li, X. Wu, Improving crowdsourced label quality us-

ing noise correction, IEEE transactions on neural networks and learning systems 29 (5) (2018) 1675–1688.

[16] L. Zhao, G. Sukthankar, R. Sukthankar, Incremental relabeling for active learning with noisy crowdsourced annotations, in: 2011 IEEE Third International Conference on Privacy, Security, Risk and Trust (PASSAT) and 2011 IEEE Third Inernational Conference on Social Computing (SocialCom), IEEE, Boston, USA, 2011, pp. 728–733. doi:10.1109/PASSAT/SocialCom.2011.193.
URL https://doi.org/10.1109/PASSAT/SocialCom.2011.193

[17] Y. Zheng, R. Cheng, S. Maniu, L. Mo, On optimality of jury selection in crowdsourcing, in: Proceedings of the 18th International Conference on Extending Database Technology, EDBT 2015, OpenProceedings. org., 2015.

[18] Y. Zheng, G. Li, Y. Li, C. Shan, R. Cheng, Truth inference in crowdsourcing: Is the problem solved?, Proceedings of the VLDB Endowment 10 (5) (2017) 541–552.

[19] A. Brew, D. Greene, P. Cunningham, Using Crowdsourcing and Active Learning to Track Sentiment in Online Media, in: Proceedings of the 19th European Conference on Artificial Intelligence (ECAI 2010), IOS Press, Lisbon, Portugal, 2010, pp. 145–150. doi:10.3233/978-1-60750-606-5-145.
URL https://doi.org/10.3233/978-1-60750-606-5-145

[20] J. Costa, C. Silva, M. Antunes, B. Ribeiro, On using crowdsourcing and active learning to improve classification performance, in: Proceedings of the 11th International Conference on Intelligent Systems Design and Applications (ISDA'11), IEEE, Córdoba, Spain, 2011, pp. 469–474. doi:10.1109/ISDA.2011.6121700.
URL https://doi.org/10.1109/ISDA.2011.6121700

[21] M. Lease, On Quality Control and Machine Learning in Crowdsourcing, in: Proceedings of the 3rd Human Computation Workshop (HCOMP) at AAAI, Association for the Advancement of Artificial Intelligence, San Francisco, USA, 2011, pp. 97–102.

[22] P. Welinder, P. Perona, Online crowdsourcing: rating annotators and obtaining cost-effective labels, in: Proceedings of the 23th Conference on Computer Vision and Pattern Recognition Workshops, IEEE, San Francisco, USA, 2010, pp. 25–32. doi:10.1109/CVPRW.2010.5543189.
URL https://doi.org/10.1109/CVPRW.2010.5543189

[23] S. Novotney, C. Callison-Burch, Cheap, Fast and Good Enough: Automatic Speech Recognition with Non-Expert Transcription, in: Proceedings of the 11th Annual Conference of the North American Chapter of the Association for Computational Linguistics, Association for Computational Linguistics, Los Angeles, USA, 2010, pp. 207–215.

[24] R. Snow, B. O'Connor, D. Jurafsky, A. Y. Ng, Cheap and fast—but is it good?: evaluating non-expert annotations for natural language tasks, in: Proceedings of the 2008 conference on Empirical Methods in Natural Language Processing, Association for Computational Linguistics, Honolulu, Hawai, 2008, pp. 254–263.

[25] J. B. Vuurens, A. P. De Vries, Obtaining high-quality relevance judgments using crowdsourcing, Internet Computing, IEEE 16 (5) (2012) 20–27. doi:10.1109/MIC.2012.71.
URL https://doi.org/10.1109/MIC.2012.71

[26] B. Frénay, M. Verleysen, Classification in the presence of label noise: a survey, IEEE Transactions on Neural Networks and Learning Systems 25 (5) (2014) 845–869. doi:10.1109/TNNLS.2013.2292894.
URL https://doi.org/10.1109/TNNLS.2013.2292894

[27] Y. Koren, R. Bell, C. Volinsky, Matrix factorization techniques for recommender systems, Computer 42 (8) (2009) 30–37. doi:10.1109/MC.2009.263.
URL https://doi.org/10.1109/MC.2009.263

[28] C. Gomes, D. Schneider, K. Moraes, J. de Souza, Crowdsourcing for music: Survey and taxonomy, in: 2012 IEEE International Conference on Systems, Man, and Cybernetic (SMC), IEEE, Seoul, Korea, 2012, pp. 832–839.

[29] A. Sorokin, D. Forsyth, Utility data annotation with amazon mechanical turk, in: Proceedings of the 21th Conference on Computer Vision and Pattern Recognition Workshops, IEEE, Anchorage, USA, 2008, pp. 1–8. doi:10.1109/CVPRW.2008.4562953.
URL https://doi.org/10.1109/CVPRW.2008.4562953

[30] M. Yuen, I. King, K. Leung, A survey of crowdsourcing systems, in: 2011 IEEE Third International Conference on Privacy, Security, Risk and Trust (PASSAT) and 2011 IEEE Third Inernational Conference on Social Computing (SocialCom), IEEE, Boston, USA, 2011, pp. 766–773. doi:10.1109/PASSAT/SocialCom.2011.203.
URL https://doi.org/10.1109/PASSAT/SocialCom.2011.203

[31] S. Buzaglo, E. Yaakobi, T. Etzion, J. Bruck, Error-correcting codes for multipermutations, in: Information Theory Proceedings (ISIT), 2013 IEEE International Symposium on, IEEE, 2013, pp. 724–728.

[32] S. Buzaglo, E. Yaakobi, T. Etzion, J. Bruck, Systematic codes for rank modulation, in: 2014 IEEE International Symposium on Information Theory (ISIT), IEEE, Honolulu, Hawai, 2014, pp. 2386–2390. doi:10.1109/ISIT.2014.6875261.
URL https://doi.org/10.1109/ISIT.2014.6875261

[33] J. Hernández-Orallo, P. Flach, C. Ferri, Roc curves in cost space, Machine Learning 93 (1) (2013) 71–91.

[34] M. G. Kendall, A new measure of rank correlation, Biometrika 30 (1/2) (1938) 81–93. doi:10.2307/2332226.
URL https://doi.org/10.2307/2332226

[35] R. Akbani, S. Kwek, N. Japkowicz, Applying support vector machines to imbalanced datasets, in: Proceedings of the 15th European Conference on Machine Learning: ECML 2004, Springer, Pisa, Italy, 2004, pp. 39–50. doi:10.1007/978-3-540-30115-8_7.
URL https://doi.org/10.1007/978-3-540-30115-8_7

[36] T. Fawcett, An introduction to ROC analysis, Pattern recognition letters 27 (8) (2006) 861–874. doi:10.1016/j.patrec.2005.10.010.
URL https://doi.org/10.1016/j.patrec.2005.10.010

[37] H. He, E. Garcia, Learning from imbalanced data, IEEE Transactions on Knowledge and Data Engineering 21 (9) (2009) 1263–1284. doi:10.1109/TKDE.2008.239.
URL https://doi.org/10.1109/TKDE.2008.239

[38] A. P. Bradley, The use of the area under the roc curve in the evaluation of machine learning algorithms, Pattern recognition 30 (7) (1997) 1145–1159. doi:10.1016/S0031-3203(96)00142-2.
URL https://doi.org/10.1016/S0031-3203(96)00142-2

[39] M. Sokolova, N. Japkowicz, S. Szpakowicz, Beyond accuracy, f-score and roc: a family of discriminant measures for performance evaluation, in: Australasian joint conference on artificial intelligence, Springer, 2006, pp. 1015–1021.

[40] T. A. Lasko, J. G. Bhagwat, K. H. Zou, L. Ohno-Machado, The use of receiver operating characteristic curves in biomedical informatics, Journal of biomedical informatics 38 (5) (2005) 404–415.

[41] M. Kubat, S. Matwin, Addressing the curse of imbalanced training sets: one-sided selection, in: Proceedings of the 14th International Conference on Machine Learning, ICML97, International Machine Learning Society, Nashville, USA, 1997, pp. 179–186.

[42] J. A. Hanley, B. J. McNeil, The meaning and use of the area under a receiver operating characteristic (ROC) curve, Radiology 143 (1) (1982) 29–36. doi:10.1148/radiology.143.1.7063747.
URL https://doi.org/10.1148/radiology.143.1.7063747

[43] J. Huang, C. X. Ling, Using AUC and accuracy in evaluating learning algorithms, IEEE Transactions on Knowledge and Data Engineering 17 (3) (2005) 299–310.

doi:10.1109/TKDE.2005.50.
URL https://doi.org/10.1109/TKDE.2005.50

[44] C. X. Ling, J. Huang, H. Zhang, Auc: a statistically consistent and more discriminating measure than accuracy, in: Proceedings of the 18th International Joint Conference on Artificial Intelligence (IJCAI), Vol. 3, IJCAI, Acapulco, Mexico, 2003, pp. 519–524.

[45] B. Satyanarayana, K. A. Mohamad, I. F. Idris, M.-L. Husain, F. Dahdouh-Guebas, Assessment of mangrove vegetation based on remote sensing and ground-truth measurements at tumpat, kelantan delta, east coast of peninsular malaysia, International Journal of Remote Sensing 32 (6) (2011) 1635–1650.

[46] P. Zhang, W. Cao, Z. Obradovic, Learning by aggregating experts and filtering novices: a solution to crowdsourcing problems in bioinformatics, BMC bioinformatics 14 (12) (2013) S5.

[47] G. Kazai, J. Kamps, M. Koolen, N. Milic-Frayling, Crowdsourcing for book search evaluation: impact of hit design on comparative system ranking, in: Proceedings of the 34th international ACM SIGIR conference on Research and development in Information Retrieval, ACM, 2011, pp. 205–214.

[48] M. Sordo, O. Celma, M. Blech, E. Guaus, The quest for musical genres: Do the experts and the wisdom of crowds agree?, in: ISMIR, 2008, pp. 255–260.

[49] R. Khatami, G. Mountrakis, S. V. Stehman, A meta-analysis of remote sensing research on supervised pixel-based land-cover image classification processes: General guidelines for practitioners and future research, Remote Sensing of Environment 177 (2016) 89–100.

[50] G. Shaw, D. Manolakis, Signal processing for hyperspectral image exploitation, IEEE Signal processing magazine 19 (1) (2002) 12–16.

[51] M. Xu, P. Watanachaturaporn, P. K. Varshney, M. K. Arora, Decision tree regression for soft classification of remote sensing data, Remote Sensing of Environment 97 (3) (2005) 322–336.

[52] H. Jeffreys, An invariant form for the prior probability in estimation problems, Proceedings of the Royal Society of London A: Mathematical, Physical and Engineering Sciences 186 (1007) (1946) 453–461. doi:10.1098/rspa.1946.0056.
URL https://doi.org/10.1098/rspa.1946.0056

[53] W. Perks, Some observations on inverse probability including a new indifference rule, Journal of the Institute of Actuaries 73 (2) (1947) 285–334.

[54] D. J. Hand, R. J. Till, A simple generalisation of the area under the roc curve for multiple class classification problems, Machine learning 45 (2) (2001) 171–186. doi:10.1023/A:1010920819831.
URL https://doi.org/10.1023/A:1010920819831

[55] C. Ferri, J. Hernández-Orallo, M. A. Salido, Volume under the roc surface for multi-class problems, in: European Conference on Machine Learning, Springer, 2003, pp. 108–120.