A note on the behavior of majority voting in multi-class domains with biased annotators

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

[1]University of the Basque Country UPV/EHU, Spain. E-mails: {jeronimo.hernandez, inaki.inza, ja.lozano}@ehu.eus
[2]Basque Center for Applied Mathematics, Spain.

Abstract

Majority voting is a popular and robust strategy to aggregate different opinions in learning from crowds, where each worker labels examples according to their own criteria. Although it has been extensively studied in the binary case, its behavior with multiple classes is not completely clear, specifically when annotations are biased. This paper attempts to fill that gap. The behavior of the majority voting strategy is studied in-depth in multi-class domains, emphasizing the effect of annotation bias. By means of a complete experimental setting, we show the limitations of the standard majority voting strategy. The use of three simple techniques that infer global information from the annotations and annotators allows us to put the performance of the majority voting strategy in context.

Index terms— Multi-class learning, Learning from crowds, Biased annotations

1 Introduction

In supervised classification, a classification model is learnt from a set of labeled examples of a specific domain so that it classifies new unlabeled examples as accurately as possible. However, obtaining the class label associated to each example for model training is usually difficult and costly.

Among other recent proposals which focus on learning with partial class information [1], learning from crowds [2] obtains (partial) class information from a crowd of workers. Workers, a.k.a. labelers, are provided with individual examples and asked to return the class label which, according to their opinion, each example belongs to. The domain knowledge of labelers may be reduced and their answer, therefore, noisy. This paradigm has received considerable attention [3] and, with the underlying assumption that mistakes are context-dependent, there already exist well-established methodologies, such as

those based on the Expectation-Maximization strategy [4, 2], which take into account the predictive variables to simultaneously infer the labeling and learn a classification model. Other techniques work only with the annotations in a step preceding the learning stage to produce a full labeling for the training examples. Consequently, it can be used to learn any classifier through standard learning techniques. In this study, we focus on these techniques and, among them, on the majority voting (MV) strategy, which stands out because of its simplicity: using the label selected by a majority of labelers. Its robust behavior under standard conditions has been extensively depicted [5, 6, 7].

Furthermore, most of the state-of-the-art studies work on binary classification [8, 9, 10]. Others [11, 2, 7] claim that their crowd learning approach straightforwardly extends to deal with many labels without going deeply into the genuine issues of the multi-class setting. While in binary classification the labeling noise just mixes two classes up, with m possible labels a labeler has $\binom{m}{2}$ ways of confusion. In this paper, we specifically explore the difficulties of MV to deal with multi-class domains when annotations are biased. In this context, bias, or recurrent noisy labeling, is defined as the trend to assign label b when the real one is c. By drawing different scenarios —such a repetitive failure may be a whole-crowd behavior or specific to certain labelers—, we aim to describe how bias affects the MV strategy. The contributions of this paper are: (i) a study of the behavior of the MV strategy in multi-class domains with biased annotations, and (ii) an empirical study on real crowd datasets.

The paper continues with a formal description of the problem. Next, our analysis of the majority voting strategy is presented, supported by different empirical studies carried out on synthetic and real crowd data. Finally, conclusions are drawn from a general discussion.

2 Multi-class learning from crowds

Formally, a supervised classification problem [12] is described by a set of predictive variables $\boldsymbol{X} = (X_1, \ldots, X_v)$ and a class variable C. Each problem example is an instance $(\boldsymbol{x}, c)$ of the random vector $(\boldsymbol{X}, C)$. Given a set of examples $D = \{(\boldsymbol{x}^1, c^1), \ldots, (\boldsymbol{x}^n, c^n)\}$, a classifier is learnt. A competitive classifier is able to generalize from D and, given a new unlabeled example, $(\boldsymbol{x}, ?)$, predict its class value, c.

Usually, a domain expert provides, from a set of possible values $\mathcal{C}$, the class value c^j associated to each training example $\boldsymbol{x}^j$. Throughout the rest of the paper, "class label" and "category" are interchangeably used to refer to any of the $m = |\mathcal{C}|$ possible values of the class variable.

Learning from crowds [2] considers a training dataset without expert supervision. By contrast, a set of noisy labelers annotates each example: $D = \{(\boldsymbol{x}^1, \boldsymbol{a}^1), \ldots, (\boldsymbol{x}^n, \boldsymbol{a}^n)\}$, where $\boldsymbol{a}^j$ is a t-tuple with $a_l^j \in \mathcal{C}$ indicating the class label assessed by labeler L_l for $\boldsymbol{x}^j$. Although ultimately the objective is also to learn a classifier, in this paper we study different approaches that estimate the real c^j from $\boldsymbol{a}^j$ in a pre-process step previous to model learning, and disregard

the descriptive information $\boldsymbol{x}^j$. Note, therefore, that no learning technique is considered in this work.

3 Majority voting in multi-class domains

The most-voted label strategy consists of selecting the category that receives the largest number of votes. When only $m = 2$ options are possible, this is equivalent to the majority voting strategy —i.e., the choice of more than a half of the voters. Although in multi-class learning ($m > 2$) these are not, strictly speaking, equivalent strategies, throughout this paper and with a little abuse of language the term "majority voting" will be used to refer to the most-voted label strategy:

$$\mathrm{MV}(\{a_1, \ldots, a_t\}) = \operatorname*{argmax}_{c \in \mathcal{C}} \sum_{l=1}^{t} \mathbb{I}[a_l = c] \tag{1}$$

where $\mathbb{I}[cond]$ is the indicator function which returns 1 if *cond* is true and 0 otherwise. The probability of the MV label being the real label c^* can be expressed as follows,

$$p(c_{\mathrm{MV}} = c^* | \boldsymbol{r}) = \sum_{\substack{(o_1,\ldots,o_m): \\ o_{c^*} \geq o_c, \forall c}} \frac{(\sum_{c=1}^{m} o_c)!}{\prod_{c=1}^{m} o_c!} \frac{\prod_{c=1}^{m} r_c^{o_c}}{\sum_{c=1}^{m} \mathbb{I}[o_c = o_{c^*}]} \tag{2}$$

where all the t labelers share the same probability distribution $\boldsymbol{r}$ (where r_c is the probability of annotating label c), and $\boldsymbol{o} = (o_1, \ldots, o_m)$ is a tuple which counts, for each class label c, the number of votes, $o_c = \sum_{l=1}^{t} \mathbb{I}[a_l = c]$.

MV is a simple yet efficient strategy with a robust behavior which has been largely studied. Its performance is enhanced as the number of annotators per example and their reliability is increased. Random mistakes are usually assumed although, if annotators tend to confuse systematically a pair of labels (i.e., label b is usually annotated when the real label is c^*), the performance of MV is compromised. Consider, for example, a domain with a *normal* category and a few more which require high expertise to identify the examples that belong to them. Intuition tells us that labelers may overpopulate the *normal* category. In this case, the MV label might not be the real one. According to Figure 1, where the effect of a biased crowd on different scenarios is depicted, annotation bias largely impacts, as expected, the performance of the MV strategy; the larger the bias, the lower the probability of MV being successful. Similarly, the impact of bias is higher when the mean reliability of the annotators (i.e., the probability of the real label, r_{c*}) decreases. Moreover, the classical trick of consulting more annotators for enhancing MV has a bare effect when the bias is large. Finally, the influence of bias seems to be reduced with large numbers of possible labels, m. Note that, in this scenario, the probability of choosing label c', $r_{c'} = \frac{1 - r_{c^*} - r_b}{m-2}$, is defined based on m, the reliability r_{c*} and the probability of bias r_b. Intuitively, given fixed values for r_{c*} and r_b, the probability of many annotators mistakenly choosing the same label c' decreases as m is enlarged

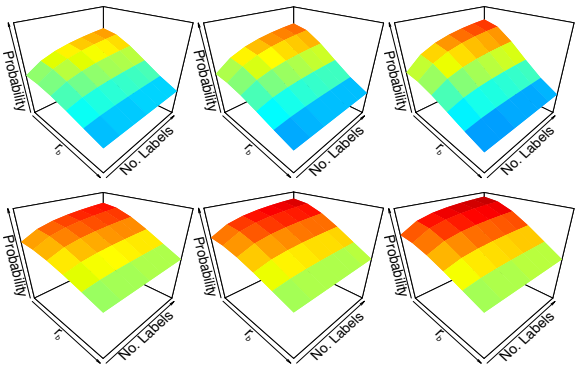

Figure 1: Probability of success of MV (Eq. 2) when annotators tend to confuse the real category, c^*, with another label, b. Each figure shows the probability as the number of labels, $m = \{3, \ldots, 9\}$, and the annotation bias, $r_b = \{\frac{1-r_{c^*}}{m-1}, \ldots, 1 - r_{c^*}\}$, are increased. Figures are displayed by column, depending on the number of annotators, $t = \{8, 10, 12\}$, and by row, depending on their mean reliability, $r_{c^*} = \{0.4, 0.5\}$.

and, consequently, the probability of success of MV would increase. As this situation might not concur in reality, similar figures are displayed in Figure 2 by ensuring a constant difference $diff = r_{c*} - r_{c'}$ between the real class and any other label (besides the biased one, b). Apart from the effect of the increase of m, which no longer benefits MV, similar behaviors are observed in the different scenarios. Again, a larger number of annotators, t, does not always allow MV to overcome the effect of bias; it is only effective when the bias is low and the difference between r_{c*} and $r_{c'}$ is large.

The MV strategy only uses the example annotations for making a decision. Information about the whole dataset is indispensable to identify bias. A simple strategy consists of selecting the label which receives the largest proportion of votes in comparison to the mean proportions of votes received. This alternative *maximum distance* (MD) strategy, is expected to be more robust than MV against biased annotations. It should detect general unbalanced distributions generated by biased annotators that noisily overpopulate a category. Depending on this, the decision of MD for each label requires a different number of votes. It is defined as,

$$\text{MD}(\boldsymbol{q}, \bar{\boldsymbol{q}}) = \underset{c \in \mathcal{C}}{\text{argmax}}(q_c - \bar{q}_c) \tag{3}$$

where $\boldsymbol{q} = (q_1, \ldots, q_m)$ is a tuple that stores, for each class label c, the proportion of votes $q_c = o_c / \sum_c o_c = \sum_l \mathbb{I}[a_l = c]/t$, and $\bar{\boldsymbol{q}}$ is a tuple with the average proportions over the whole training dataset, $\bar{q}_c = \sum_{j=1}^{n} q_c^j / n$. However, a gain of 1 vote might not have the same relevance in a category which usually obtains, for instance, 100 votes than a gain of an extra vote for a category which usually

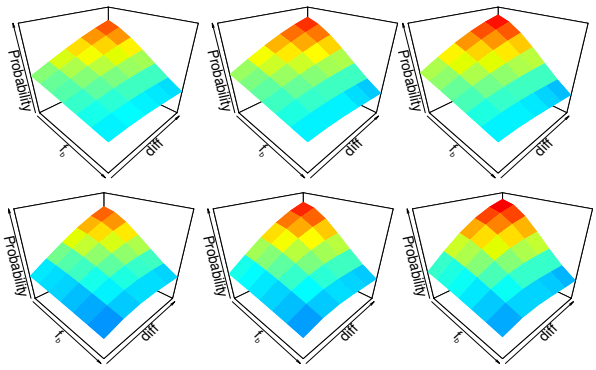

Figure 2: Probability of success of MV (Eq. 2) when annotators tend to confuse the real category, c^*, with another label, b. A fixed difference, $diff$, is guaranteed between r_{c*} and $r_{c'}$, $\forall c' \neq c* \neq b$. Each figure shows the probability as the difference, $diff$, and the bias, $r_b = f_b \cdot r_{c^*}$ (where $f_b = \{0, 0.25, \ldots, 1.5\}$), are increased. Figures are displayed by column, depending on the number of annotators, $t = \{8, 10, 12\}$, and by row, depending on the number of labels, $m = \{3, 9\}$.

Table 1: Used datasets [13, 14] described by no. examples (n), no. labels (m) and no. examples per label, ordered by imbalance degree (ID_{HE}) [15].

Dataset	n	m	Label distribution	ID_{HE}
vowel	990	11	{90×11}	0.000
segment	2310	7	{330×7}	0.000
vehicle	846	4	{212, 217, 218, 199}	0.008
svmguide4	612	6	{86, 116, 119, 99, 110, 82 }	0.348
satimage	6435	6	{1533, 703, 1358, 626, 707, 1508}	0.372
dermatology	366	6	{112, 61, 72, 49, 52, 20}	0.381
pendigits	10992	10	{1142, 1143×2, 1144×2, 1055×4, 1056}	0.402
glass	214	6	{70, 76, 17, 13, 9, 29}	0.567
usps	9298	10	{1553, 1269, 929, 824, 852, 716, 834, 792, 708, 821}	0.712
arrhythmia	452	13	{245, 44, 15, 15, 13, 25, 3, 2, 9, 50, 4, 5, 22}	0.738

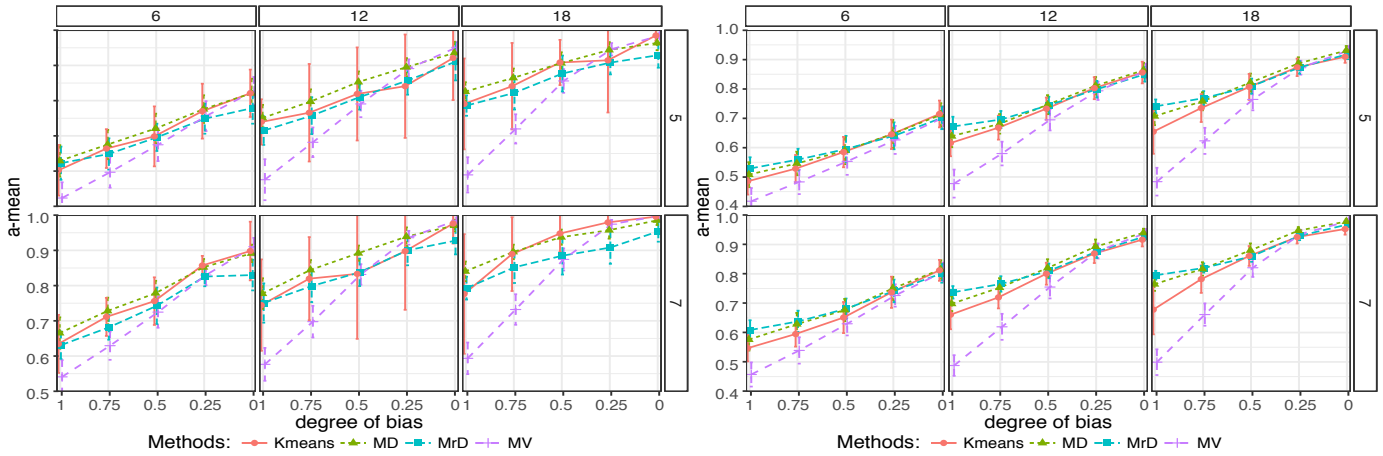

Figure 3: Results of the four aggregation functions in terms of a-mean and associated standard deviation. In the left figure, synthetic datasets are used ($m = 5$) and, in the right figure, real datasets (Tab. 1). In both figures, plots are displayed by column, depending on the number of annotators, $t = \{6, 12, 18\}$, and by row, depending on the *relevance*, $\{5, 7\}$, of the real label. Each plot shows performance as the bias degree (α) is reduced.

gets 2. This is attained by aggregations based on relative distance, such as,

$$\mathrm{MrD}(\boldsymbol{q}, \bar{\boldsymbol{q}}) = \underset{c \in \mathcal{C}}{\operatorname{argmax}} \, (q_c / \bar{q}_c) \tag{4}$$

where MrD stands for *maximum relative distance*. The supplementary material, at the website associated with this paper[1], includes pseudo-codes for these methods, the used source codes, and a simple example to illustrate how they work.

3.1 Experimental comparison on biased crowds

With the objective of analyzing the behavior of the MV strategy regarding other distribution-aware strategies, a set of experiments has been carried out in synthetically biased data. Specifically, different scenarios where all the labelers undergo the same kind of bias have been simulated.

Both real multi-class problems (selected from public repositories [13, 14], see Table 1) and synthetic data (generated as explained in the Appendix, with $m = 5$) are used.

As also explained in the Appendix, a labeler is modeled by a probability distribution over the labels, and their annotations are generated by sampling it. Different types of annotator are simulated using three intelligible parameters. Firstly, the reliability, i.e., the probability of annotating the right label

[1] `http://www.sc.ehu.es/ccwbayes/members/jeronimo/crbias/`

c^*, is determined by the *relevance* parameter. On average, a higher *relevance* corresponds to labelers with higher reliability. Secondly, γ controls the rate of annotators with a biased behavior in a crowd. Biased annotators recurrently annotate label b (fixed) when the real category is c^*. Finally, the bias degree, $\alpha \in [0, 1]$, determines the strength of the bias, from the most biased labelers ($\alpha \to 1$) to the least ($\alpha = 0$). In this set of experiments, sets of annotators are simulated with parameters $\gamma = 1$, two *relevance* values ($\{5, 7\}$) and the whole spectrum of bias degree ($\alpha = \{0, \ldots, 1\}$). In each experiment, a single distribution is used to simulate t labelers and their behavior in all the categories.

Experiments are validated using the a-mean metric, i.e., the average recall over the m classes,

$$\text{a-mean}(\boldsymbol{h}, \hat{\boldsymbol{h}}) = \frac{1}{m} \sum_{c=1}^{m} \frac{\sum_{j=1}^{n} \mathbb{I}[(h^j = c) \wedge (\hat{h}^j = c)]}{\sum_{j=1}^{n} \mathbb{I}[h^j = c]} \tag{5}$$

where $\boldsymbol{h}$ is the real labeling and $\hat{\boldsymbol{h}}$ is an aggregated one. This metric equally values the performance in all the class labels. This fact distinguishes a-mean from global-performance metrics (where competent performance in dominant classes can mask an extremely poor behavior in underrepresented categories), such as accuracy, and makes it especially suitable when the objective is to perform robustly across all the labels [15]. In this case, an across-label robust performance is necessary to obtain labeled examples of all the classes for the subsequent training stage. Results in terms of other popular (im)balance-aware metrics, AUC [16] and F1 [17], are shown in the supplementary material at the associated website.

In addition to the three simple techniques, a method [18] recently proposed to deal with biased annotations is also considered to broaden the experimental spectrum. The k-means clustering algorithm with $k = m$ (number of class labels) is run on a dataset formed by the proportion of votes, $\boldsymbol{q}$, together with the mean leap length between consecutive categories ($q_c - q_{c-1}$), as variables. The cluster with the centroid $\hat{\boldsymbol{q}}$ with the largest value $\hat{q}_c$, and all the examples in it, is assigned to class label, c. As labels are assigned based on the proportions of annotations $\boldsymbol{q}$, and not based on a single value, this method can be considered as bias-robust.

In Figure 3, the performance of MV, MD, MrD and k-means based approaches is displayed. The behavior is similar in both real and synthetic data. Bias harms all the approaches, although the MV strategy is consistently the most affected one. It is only competitive with the rest of the approaches when the bias is slight ($\alpha = 0.25$) or null ($\alpha = 0$). The number of annotators, t, has a limited effect on MV when the bias is large (note the steeper slope as t grows), whereas a performance improvement is observed as t grows (as usual in the field) for the rest of aggregation functions. The *relevance* of the real class label does not have a predominant effect. The main challenge seems to be the difference among the probability of the real class r_{c*} and the bias r_b, and not the difference with the rest of categories.

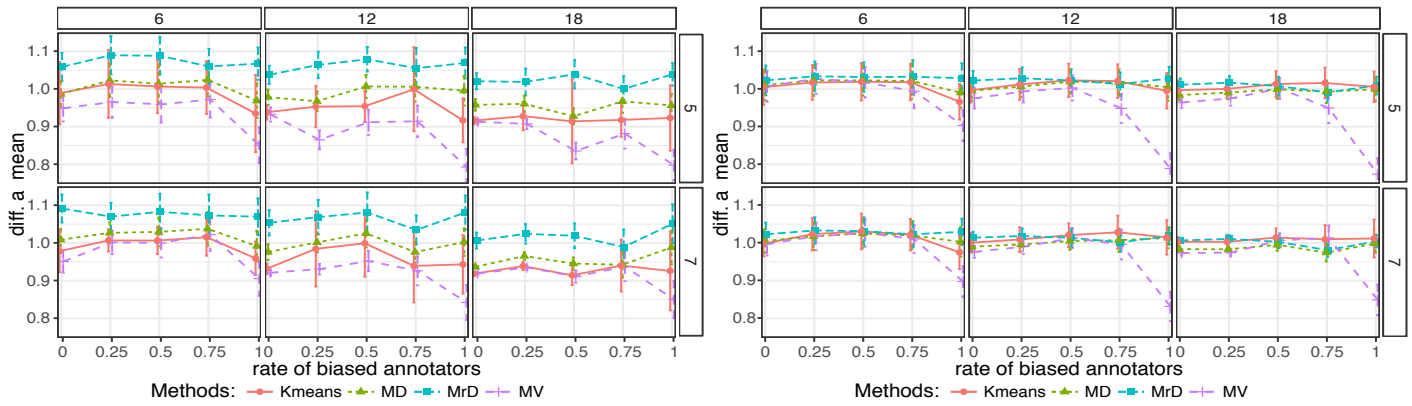

Figure 4: Proportional difference in terms of a-mean of the results of wMV in combination with the four aggregation functions regarding the use of the four aggregators alone. In the left figure, synthetic datasets are used ($m = 5$) and, in the right figure, real datasets (Tab. 1). In both figures, plots are displayed by column, depending on the number of annotators, $t = \{6, 12, 18\}$, and by row, depending on the *relevance*, $\{5, 7\}$, of the real label. The bias degree is fixed ($\alpha = 0.75$). Each plot shows the performance difference as the rate of biased annotators (γ) increases: A value larger than 1 in the y-axis depicts a scenario where the use of the aggregation function for weight estimation outperforms the use of the same aggregator alone.

Questioned by the possible influence of the issue of class imbalance on these results, the averaged results of Figure 3 were expanded in the spectrum of the different imbalance degrees [15] shown by the datasets of Table 1 (available in the supplementary material at the associated website). Using a moving average of length 3, it can be seen that class imbalance affects all the approaches similarly. That is, it has no observable influence on the results of these experiments.

Among the rest of approaches, the more elaborate k-means based strategy does not outperform the simple MD strategy. The prevalence of MD is clearer as the degree of bias is increased ($\alpha \to 1$). These results recall the robustness of simple approaches in the aggregation of multiple annotations. Although in real data MD and MrD perform similarly, with synthetic datasets MD clearly outperforms MrD.

4 Weighted majority voting

Although not all the annotators show a biased annotation behavior, a subset of highly biased annotators may be enough to infer deteriorated labelings. In this case, one might be tempted to model who is right and when. A common strategy is to use a set of parameters to model the reliability of the labelers and accordingly count their votes in a *weighted majority voting* (wMV) approach:

$$\text{wMV}(\boldsymbol{a}, \boldsymbol{w}) = \underset{c \in \mathcal{C}}{\text{argmax}} \sum_{l=1}^{t} w_c^l \cdot \mathbb{I}[a_l = c] \tag{6}$$

where w_c^l is the reliability of labeler L_l with label c. wMV requires the estimation of reliability weights, which is ideally carried out by comparing the annotations of each labeler with the ground truth. In the lack of real labels, weight estimation needs to rely on an approximated ground truth,

$$w_c^l = \begin{cases} \frac{\sum_{j=1}^{n} \mathbb{I}[(agg(\boldsymbol{a}^j)=c) \wedge (a_l^j=c)]}{\sum_{j=1}^{n} \mathbb{I}[a_l^j=c]} & \text{if } \sum_{j=1}^{n} \mathbb{I}[a_l^j = c] > 0 \\ 0 & \text{otherwise} \end{cases} \tag{7}$$

where $agg(\cdot)$ is an aggregation function. Note that the quality of the estimated weights, w_c^l, depends on how the aggregated labeling conforms with reality. Usually, the aggregation function is MV ($agg(\cdot) \equiv$ Eq. 1), and w_c^l is, therefore, the probability that the MV label is c when labeler L_l suggests c. However, throughout this paper we have seen that MV is not robust against bias. Alternative aggregation functions, $agg(\cdot)$, could be used to estimate the weights w_c^l.

4.1 Experiments on crowds with a few biased labelers

In these experiments, MD, MrD and k-means based approach are used, in addition to MV, as aggregation functions for estimating the weights for wMV in Equation 7. Similar to the previous experimental setting, both real (Tab. 1) and synthetic data (with $m = 5$, see Appendix) is used. Annotations are generated

with *relevance* values of $\{5, 7\}$ (see Appendix). Biased (with $\alpha = 0.75$) and unbiased annotators are generated in different proportions, $\gamma = \{0, \ldots, 1\}$.

Figure 4 depicts the behavior of wMV in combination with the four aggregation functions regarding the use of the aggregators alone. In different scenarios, each subfigure shows the rate —(wMV+agg)/agg— for each aggregation function (agg) using a-mean (Eq. 5): A value larger than 1 in the y-axis represents a scenario where the use of agg for weight estimation outperforms the use of agg alone. Results using AUC and F1 metrics are available in the supplementary material at the associated website. According to these results, MV seems inadequate to estimate weights, even when no biased annotator is present in the crowd ($\gamma = 0$). In almost all the scenarios, the MV strategy works better than the wMV+MV combination. As observed for the four aggregation functions, the usefulness of a wMV strategy decays as the number of annotators (t) is increased. On the contrary, a larger *relevance* of the annotator reliability on the real label enhances the performance of the wMV alternatives. Moreover, when the proportion of biased annotators is large ($\gamma > 0.5$), the relative performance of the weighted approaches drops. From $\gamma = 0$ to 0.5, the performance slightly improves with the proportion of biased labelers. Somehow, the presence of a few biased annotators eases the identification of the individual biases, but properly estimating the weights of the annotators is unfeasible when the mistaken behavior is widespread. Similar behaviors are observed in synthetic and real datasets, although the differences are much slighter in the latter. The stability of experiments with real datasets (i.e., the straightness of the lines of results) is also noteworthy. This may be a product of the class cardinality, m, which in synthetic data is constant and somehow low regarding the cardinality of the real datasets.

According to Fig. 4, the use of weights consistently improves the MrD strategy. On the contrary, the improvement of the rest of strategies when combined with wMV is generally slight and mainly limited to experiments with reduced number of annotators, t, and medium rates of biased labelers, γ. That is, the integration does not work when the scenario is simple enough for the original aggregation function alone, neither does it work in extremely complex scenarios where both approaches perform poorly. The issue of class imbalance has also been analyzed by means of a moving average over the datasets ordered by imbalance degree [15] (see the supplementary material at the associated website). The weighted approach works robustly with MrD in highly unbalanced datasets. Apparently, other techniques are only affected by this issue when a large proportion of annotators are biased. This delicate issue is an open question which should be analyzed more deeply in a dedicated study.

Note that our decision on the bias degree ($\alpha = 0.75$) determines the results. A larger bias would enhance the performance of wMV approaches and, consistently, aggregation functions alone perform better as the bias degree is reduced.

Table 2: Used real crowd datasets [3] described by number of examples (n), number of labels (m), number of examples per label, number of labelers (t), and number of annotations per example and per labeler.

Dataset	n	m	Label distribution	$\mathrm{ID}_{\mathrm{HE}}$	t	$\frac{\#ann}{n}$	$\frac{\#ann}{t}$
music_genre	700	10	{67, 72×3, 74, 63, 71, 75, 64, 70}	0.204	44	66.9	4.2
valence5	100	5	$\{13, 27, 23, 29, 8\}$	0.262	38	10.0	26.3
dogs	800	4	$\{169, 185, 218, 228\}$	0.269	109	10.0	73.4
saj2013	300	5	$\{57, 70, 72, 92, 9\}$	0.284	461	5.7	3.7
wordsim5	30	5	$\{1, 12, 4, 6, 7\}$	0.294	10	10.0	30.0
trec2010	3275	3	$\{1500, 863, 912\}$	0.380	722	5.6	25.6
weather_sent	291	4	$\{57, 70, 72, 92\}$	0.521	110	19.1	50.5
fej2013	576	3	$\{19, 531, 26\}$	0.573	48	5.0	60.5
adult2	333	4	$\{187, 61, 36, 49\}$	0.583	269	9.9	12.3

5 Discussion

The performance of the MV strategy does not reach that of the other studied strategies when the bias is relevant. Contrary to general belief, its performance in these circumstances is not significantly improved with more labelers. The limited performance of MV in the presence of a considerable bias also implies that its aggregations are not useful for weight estimation. The k-means based approach [18], specifically designed to work with biased annotations, is a solid aggregator which establishes a competitive baseline. Nevertheless, the rest of the simpler techniques show a similar performance and, commonly, overcome this specifically designed approach. The approaches based on maximum distance stand out for both their simplicity and their performance in a context of biased labelers. Whereas the MD strategy is very competitive when it works alone, MrD is the aggregation function which most improves wMV.

Our final experiments on real multi-class crowd data (selected from a public repository [3], see Table 2) support these considerations. The performance of the tested aggregation functions, both alone and in combination with wMV, is displayed in Table 3 in terms of a-mean. Results in terms of AUC and F1 metrics are available in the supplementary material at the associated website.

The performance of MV is again limited in comparison with the rest of strategies. In concordance with previous results, the competitive behavior of MD and MrD is notable. In fact, the k-means based approach, which is specifically designed to deal with biased datasets, does not outperform the basic approaches. However, the MV strategy is clearly outperformed by its competitors in terms of a-mean. Although not presented in the paper due to space limitations, the MV strategy is competitive with respect to the other approaches in terms of accuracy. A great leap is observed between the performance of MV in terms of

Table 3: Results in terms of a-mean of the four aggregation functions on real crowd datasets, alone and in combination with weighted voting.

Dataset	MV	MD	MrD	k-means
music_genre	0.712	0.710	0.705	**0.717**
valence5	0.371	0.516	**0.537**	0.483
dogs	0.820	**0.831**	**0.832**	0.809
saj2013	0.788	0.789	**0.806**	0.790
wordsim5	0.475	0.633	**0.700**	0.550
trec2010	0.453	**0.465**	**0.466**	0.449
weather_sent	**0.885**	**0.887**	0.880	**0.885**
fej2013	0.647	0.620	0.633	**0.662**
adult2	0.598	**0.696**	0.678	0.661
average	0.639	0.683	**0.693**	0.667
	Weighted voting + *agg*			
music_genre	**0.795**	0.783	0.781	0.789
valence5	0.314	0.482	**0.508**	0.414
dogs	0.822	**0.837**	**0.837**	0.810
saj2013	0.798	0.789	**0.809**	0.790
wordsim5	0.400	0.500	**0.633**	0.400
trec2010	0.434	**0.472**	**0.472**	0.422
weather_sent	**0.889**	**0.890**	0.887	0.885
fej2013	**0.650**	0.645	0.646	**0.651**
adult2	0.583	**0.660**	0.647	0.618
average	0.632	0.673	**0.691**	0.642

accuracy and a-mean in saj2013 and fej2013 datasets. Presumably, at least one of the class labels is disregarded by the MV strategy: in both datasets there is a minority category which is possibly never considered. However, in most cases, a single approach is the best one in terms of both a-mean and accuracy. This shows once again the robustness of basic techniques.

Based on these experimental results, it can be concluded that weighted voting is not always an appropriate strategy. It seems unnecessary, for example, when many annotators are available. In the real crowd datasets, although the performance gains may be noteworthy (music_genre, dogs or trec2010), the results are not always enhanced by the use of weighted voting (in saj2013 or weather_sent, the results are similar with and without weights, and, in other cases, weighted voting is worse). In both synthetic and real crowds, MD and MrD are the best combination for the wMV strategy to estimate the weights of the labelers. This is another indicator of the ability of the maximum distance aggregators to obtain the real label in biased domains.

6 Conclusions and future work

In this paper, we study the behavior of the majority voting strategy dealing with biased annotations. Its lack of perspective —aggregation is performed without taking into account global behaviors such as bias— limits its performance. Standard decisions such as enlarging the number of annotations are not efficient enough to compensate the effect of bias. These troubles may even worsen if MV is used in combination with a weighted approach to estimate the reliability of annotators. Other strategies specifically designed to deal with biased annotations clearly overcome MV in biased domains. Specifically, both simple approaches based on maximum distance turned out to be notably competitive.

Only simple techniques that do not consider the example descriptions ($\boldsymbol{x}$) were studied. However, when this information is available (e.g., to learn a classifier), we could take advantage of it. In this context, measuring to what extent the example descriptors can enhance the estimation of the ground truth labels would be of interest. It could be also interesting to study the robustness of the distance based approaches if annotator reliability weights are introduced directly in their calculation (Eqs. 3 and 4), in a similar way as wMV (Eq. 6) does with MV (Eq. 1). Finally, an in-depth study which analyzes the effect of class imbalance on the behavior of the annotators and its final impact on the estimated ground truth would definitely be valuable.

Acknowledgments

This work has been partially supported by the Basque Government (IT609-13, Elkartek BID3A), the Spanish Ministry of Economy and Competitiveness (TIN2016-78365-R) and the University-Society Project 15/19 (Basque Gov-

ernment and University of the Basque Country UPV/EHU). J.A. Lozano is also supported by BERC program 2014-2017 (Basque Government) and Severo Ochoa Program SEV-2013-0323 (Spanish Ministry of Economy and Competitiveness).

A synthetic domain (ground truth) is generated by sampling a Dirichlet distribution, $\mathcal{D}$, with as many hyper-parameters (all equal to 1) as class labels. A sample generates the parameters of a categorical distribution which is, in turn, sampled to obtain the artificial labeling (as many times as examples are required in the dataset). Experiments were replicated 20 times (samples of $\mathcal{D}$) and averaged among different ground truths.

To generate synthetic crowd annotations, first of all, a set of parameters $\boldsymbol{\theta}^{lc}$ is generated for each labeler L_l and label c by sampling a Dirichlet distribution, $\mathcal{B}$, with m hyper-parameters. A sample generates the parameters of a categorical distribution which is, in turn, sampled to generate the annotations. Given a labeled example $(\boldsymbol{x}^j, c^j)$, the categorical distribution $\boldsymbol{\theta}^{lc^j}$ is sampled to obtain a_l^j, the label annotated by annotator L_l. Experiments were replicated 20 times (samples of $\mathcal{B}$) and the results averaged.

By controlling the hyper-parameters $\{\beta_c\}_{c=1}^m$ of $\mathcal{B}$, different types of labelers are simulated. Hyper-parameters are established according to the *relevance*, the *rate* of biased annotators (γ) and the bias degree (α) parameters. The beta value corresponding to the real label is set as $\beta_{c^*} = \mathit{relevance} > 1$ (and $\forall c \neq c^*, \beta_c = 1$). Biased annotators use a different distribution $\mathcal{B}$ with a modified hyperparameter for label b: $\beta_b = \alpha\beta_{c^*} + (1-\alpha)\beta_c$, $\beta_c \neq \beta_{c^*}$. In both cases, hyper-parameters β_c are normalized, $\beta_0 \cdot \beta_c / \sum_{c'} \beta_{c'}$, to fix $\beta_0 = 10$.

References

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

[2] 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. 1297–322, 2010.

[3] J. Zhang, X. Wu, and V. S. Sheng, "Learning from crowdsourced labeled data: a survey," *Artif. Intell. Rev.*, vol. 46, no. 4, pp. 543–76, 2016.

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

[5] 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 *Proc. EMNLP 2008*, 2008, pp. 254–63.

[6] V. S. Sheng, F. J. Provost, and P. G. Ipeirotis, "Get another label? improving data quality and data mining using multiple, noisy labelers," in *Proc. 14th ACM SIGKDD*, 2008, pp. 614–22.

[7] J. Hernández-González, I. Inza, and J. A. Lozano, "Multidimensional learning from crowds: Usefulness and application of expertise detection," *Int. J. Intell. Syst.*, vol. 30, no. 3, pp. 326–54, 2015.

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

[9] P. Welinder, S. Branson, S. Belongie, and P. Perona, "The multidimensional wisdom of crowds," in *Proc. 23rd NIPS*, 2010.

[10] J. Zhang, X. Wu, and V. S. Sheng, "Imbalanced multiple noisy labeling," *IEEE Trans. Knowl. Data Eng.*, vol. 27, no. 2, pp. 489–503, 2015.

[11] G. Demartini, D. E. Difallah, and P. Cudré-Mauroux, "Zencrowd: leveraging probabilistic reasoning and crowdsourcing techniques for large-scale entity linking," in *Proc. 21st Int. Conf. WWW*, 2012, pp. 469–78.

[12] T. Mitchell, *Machine Learning*. McGraw Hill, 1997.

[13] C.-C. Chang and C.-J. Lin, "LIBSVM: a library for support vector machines," 2011. [Online]. Available: http://www.csie.ntu.edu.tw/~cjlin/libsvmtools/datasets/

[14] A. Frank and A. Asuncion, "UCI machine learning repository," 2010. [Online]. Available: http://archive.ics.uci.edu/ml

[15] J. Ortigosa-Hernández, I. Inza, and J. A. Lozano, "Measuring the class-imbalance extent of multi-class problems," *Pattern Recognit. Lett.*, vol. 98, pp. 32–38, 2017.

[16] M. Galar, A. Fernandez, E. Barrenechea, H. Bustince, and F. Herrera, "A review on ensembles for the class imbalance problem: Bagging-, boosting-, and hybrid-based approaches," *IEEE Trans. Syst. Man Cybern. Part C-Appl. Rev.*, vol. 42, no. 4, pp. 463–84, 2012.

[17] S. Wang and X. Yao, "Multiclass imbalance problems: Analysis and potential solutions," *IEEE Trans. Syst. Man Cybern. Part B-Cybern.*, vol. 42, no. 4, pp. 1119–30, 2012.

[18] 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.