Supplementary material:
“A note on the behavior of majority voting in multi-class domains with biased annotators”

Jeronimo Hernandez-Gonzalez, Inaki Inza, and Jose A. Lozano, Member, IEEE

Abstract

In this document, supplementary material∗ is provided for the paper entitled “A note on the behavior of majority voting in multi-class domains with biased annotators”.

- Pseudo codes are provided for all the implemented algorithms: MV, Alg. 1; MD, Alg. 2; MrD, Alg. 3; k-means based approach, Alg. 4; and wMV, Alg. 5.
- An example of the calculations of MV, MD and MrD is provided in Table 1.
- Results of the experiments displayed in Figures 3 and 4 of the main paper are displayed in terms of other metrics: (Macro) F1-measure in Figures 1 and 3; and (Macro) AUC in Figures 2 and 4.
- Two different sets of experiments, equivalent to those of Figures 3 and 4 of the main paper (and Figures 1 to 4 in this document), are displayed taking into account the issue of class imbalance in terms of a-mean (Figures 5 and 8), Macro F1 (Figures 6 and 9), and Macro AUC (Figures 7 and 10).
- Results of the experiments displayed in Table 3 of the main paper are displayed in terms of other metrics: (Macro) F1-measure in Table 2, and (Macro) AUC in Table 3.

∗ See the main paper for definition of symbols and other relevant information.

Algorithm 1 Pseudocode of the majority voting (MV) approach.

```
procedure MV({a_1, a_2, . . . , a_n})
    \hat{h} \leftarrow \text{new tuple}(n\text{Elements}:n)
    \text{for } j \in \{1, . . . , n\} \text{ do}
        q \leftarrow \text{new tuple}(n\text{Elements}:|C|)
        \text{for } c \in \{1, . . . , |C|\} \text{ do}
            q_c \leftarrow \text{countsOfLabel}(a_j, c)
        end for
        mv \leftarrow \text{which}(\{q_c = \max(q_{c-1})\})\text{|}_c
        \text{if } |mv| = 1 \text{ then}
            \hat{h}_j \leftarrow mv_1 \text{ No. annotators providing label } c \text{ in } a_j
        \text{else if } |mv| > 1 \text{ then}
            \hat{h}_j \leftarrow \text{randomSelection}(mv) \text{ Label(s) which have received the largest number of votes}
        end if
    end for
    return \{\hat{h}_1, \hat{h}_2, . . . , \hat{h}_n\}
end procedure
```
Algorithm 2 Pseudocode of the maximum distance (MD) approach.

```plaintext
procedure MD((a1, a2, ..., an))
    Q ← new matrix(nRow: n, nCol: |C|)  \(\triangleright q_{jc}\) is the cell in the intersection of the j-th row and the c-th column of Q
    for j ∈ {1, ..., n} do
        for c ∈ {1, ..., |C|} do
            q_{jc} ← countsOfLabel(a_j, c)  \(\triangleright \) No. annotators providing label c in a_j
        end for
    end for
    \(q̄\) ← meanByRow(Q)
    \(h\) ← new tuple(nElements:n)
    for j ∈ {1, ..., n} do
        \(mv\) ← which(\(q_{jc} = max(q_{jc})\))_{c=1}^{|C|}  \(\triangleright \) Label(s) which have received the largest number of votes 
        \(\triangleright \) in comparison with its mean
        if |mv| = 1 then
            \(h_j\) ← mv_1  \(\triangleright \) Each example is assigned to the label c with the largest number of votes (mv_1)
        else if |mv| > 1 then
            \(h_j\) ← randomSelection(mv)  \(\triangleright \) Ties are solved randomly: any label with the maximum number of votes
        end if
    end for
    return \{\(h_1, h_2, ..., h_n\}\}
end procedure
```

Algorithm 3 Pseudocode of the maximum relative distance (MrD) approach.

```plaintext
procedure MRD((a1, a2, ..., an))
    Q ← new matrix(nRow: n, nCol: |C|)  \(\triangleright q_{jc}\) is the cell in the intersection of the j-th row and the c-th column of Q
    for j ∈ {1, ..., n} do
        for c ∈ {1, ..., |C|} do
            q_{jc} ← countsOfLabel(a_j, c)  \(\triangleright \) No. annotators providing label c in a_j
        end for
    end for
    \(q̄\) ← meanByRow(Q)
    \(h\) ← new tuple(nElements:n)
    for j ∈ {1, ..., n} do
        \(mv\) ← which(\(q_{jc}/q̄_c = max(q_{jc}/q̄_c)\))_{c=1}^{|C|}  \(\triangleright \) Label(s) which have received the largest number of votes
        \(\triangleright \) relative to its mean
        if |mv| = 1 then
            \(h_j\) ← mv_1  \(\triangleright \) Each example is assigned to the label c with the largest number of votes (mv_1)
        else if |mv| > 1 then
            \(h_j\) ← randomSelection(mv)  \(\triangleright \) Ties are solved randomly: any label with the maximum number of votes
        end if
    end for
    return \{\(h_1, h_2, ..., h_n\}\}
end procedure
```

Algorithm 4 Pseudocode of the k-means based approach.

```plaintext
procedure K-MEANS((a1, a2, ..., an))
    Q ← new matrix(nRow: n, nCol: |C| + 1)  \(\triangleright q_{jc}\) is the cell in the intersection of the j-th row and the c-th column of Q
    for j ∈ {1, ..., n} do
        for c ∈ {1, ..., |C|} do
            q_{jc} ← countsOfLabel(a_j, c)  \(\triangleright \) No. annotators providing label c in a_j
        end for
    end for
    \(q̄_{(c+1)}\) ← \(\sum_{c=2}^{|C|} q_{jc} = q_{j(c-1)}\)  \(\triangleright \) Each centroid represents a class label c (k = |C|)
    \(C_{Centroids}\) ← \{arg max_{j∈1,...,n} q_{jc}\}_{c=1}^{|C|}
    \(h\) ← Kmeans(Q, iCentroids, k = |C|)  \(\triangleright \) Assign each example to a centroid
    return \{\(h_1, h_2, ..., h_n\}\}  \(\triangleright \) Examples grouped by k-means with the centroid of label c are assigned to label c
end procedure
```
Algorithm 5 Pseudocode of the weighted majority voting (wMV) approach.

procedure wMV(\{a_1, a_2, \ldots, a_n\})
    \(W \leftarrow \text{new matrix}(nRow, nCol | C|)\)
    \(h = \text{agg}(\{a_1, a_2, \ldots, a_n\})\)
    for \(c \in \{1, \ldots, |C|\}\) do
        \(h^c \leftarrow \text{examplesOfLabel}(h, c)\)
        for \(t \in \{1, \ldots, t\}\) do
            \(a_t^c \leftarrow \text{annotationsOfLabel}(\{a_1, a_2, \ldots, a_n\}, t, c)\)
            \(w_{c,t}^t \leftarrow |h^c \cap a_t^c|/||h^c||\)
        end for
    end for
    \(\hat{h} \leftarrow \text{new tuple(nElements:n)}\)
    for \(j \in \{1, \ldots, n\}\) do
        for \(c \in \{1, \ldots, |C|\}\) do
            \(q_{c,j} \leftarrow \text{new tuple(nElements:|C|)}\)
            \(q_{c,j} = \sum_{t=1}^{t} w_{c,t}^t \cdot [a_t^c = c]\)
        end for
        \(m_{v,j} \leftarrow \text{which}(\{q_{c,j} = \max(q_{c,j})\}_{c=1}^{|C|})\)
        if \(|m_{v,j}| = 1\) then
            \(h_j \leftarrow m_{v,j}\)
        else if \(|m_{v,j}| > 1\) then
            \(h_j \leftarrow \text{randomSelection}(m_{v,j})\)
        end if
    end for
    return \(\{h_1, h_2, \ldots, h_n\}\)
end procedure

TABLE 1
Example of the use of three aggregate functions (majority voting –MV–, maximum distance –MD– and maximum relative distance –MrD–) in an illustrative example with 3 class labels and 6 annotators. The average proportions of annotators used for the calculation of MD and MrD are \(q = \{0.54, 0.31, 0.15\}\). In case of a tie, the first class label is selected.

<table>
<thead>
<tr>
<th>Annotations</th>
<th>Proportions of annots.</th>
<th>(q - \bar{q})</th>
<th>(q/\bar{q})</th>
<th>Results</th>
</tr>
</thead>
<tbody>
<tr>
<td>(L_1)</td>
<td>(L_2)</td>
<td>(L_3)</td>
<td>(L_4)</td>
<td>(L_5)</td>
</tr>
<tr>
<td>1</td>
<td>1</td>
<td>2</td>
<td>1</td>
<td>2</td>
</tr>
<tr>
<td>1</td>
<td>1</td>
<td>2</td>
<td>2</td>
<td>1</td>
</tr>
<tr>
<td>2</td>
<td>3</td>
<td>2</td>
<td>3</td>
<td>2</td>
</tr>
<tr>
<td>3</td>
<td>2</td>
<td>1</td>
<td>2</td>
<td>2</td>
</tr>
<tr>
<td>1</td>
<td>2</td>
<td>1</td>
<td>1</td>
<td>2</td>
</tr>
<tr>
<td>3</td>
<td>2</td>
<td>3</td>
<td>3</td>
<td>2</td>
</tr>
<tr>
<td>3</td>
<td>1</td>
<td>2</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>2</td>
</tr>
<tr>
<td>3</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>3</td>
</tr>
<tr>
<td>3</td>
<td>3</td>
<td>1</td>
<td>2</td>
<td>1</td>
</tr>
<tr>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
</tr>
<tr>
<td>3</td>
<td>1</td>
<td>2</td>
<td>1</td>
<td>2</td>
</tr>
<tr>
<td>1</td>
<td>2</td>
<td>1</td>
<td>2</td>
<td>2</td>
</tr>
</tbody>
</table>

TABLE 2
Results in terms of macroF1 of the four aggregation functions on real crowd datasets, alone and in combination with weighted voting.

<table>
<thead>
<tr>
<th>Dataset</th>
<th>MV</th>
<th>MD</th>
<th>MrD</th>
<th>k-means</th>
<th>MV</th>
<th>MD</th>
<th>MrD</th>
<th>k-means</th>
</tr>
</thead>
<tbody>
<tr>
<td>adult2</td>
<td>0.636</td>
<td>0.706</td>
<td>0.674</td>
<td>0.688</td>
<td>0.627</td>
<td>0.802</td>
<td>0.669</td>
<td>0.652</td>
</tr>
<tr>
<td>dogs</td>
<td>0.823</td>
<td>0.831</td>
<td>0.833</td>
<td>0.813</td>
<td>0.836</td>
<td>0.839</td>
<td>0.841</td>
<td>0.823</td>
</tr>
<tr>
<td>fe2013</td>
<td>0.507</td>
<td>0.477</td>
<td>0.559</td>
<td>0.543</td>
<td>0.508</td>
<td>0.507</td>
<td>0.507</td>
<td>0.515</td>
</tr>
<tr>
<td>music_genre</td>
<td>0.713</td>
<td>0.709</td>
<td>0.703</td>
<td>0.722</td>
<td>0.812</td>
<td>0.789</td>
<td>0.785</td>
<td>0.815</td>
</tr>
<tr>
<td>sal2013</td>
<td>0.785</td>
<td>0.78</td>
<td>0.795</td>
<td>0.793</td>
<td>0.808</td>
<td>0.78</td>
<td>0.806</td>
<td>0.777</td>
</tr>
<tr>
<td>tree2010</td>
<td>0.461</td>
<td>0.468</td>
<td>0.469</td>
<td>0.459</td>
<td>0.462</td>
<td>0.469</td>
<td>0.469</td>
<td>0.463</td>
</tr>
<tr>
<td>valence5</td>
<td>0.416</td>
<td>0.491</td>
<td>0.514</td>
<td>0.501</td>
<td>0.221</td>
<td>0.375</td>
<td>0.495</td>
<td>0.426</td>
</tr>
<tr>
<td>wordsim5</td>
<td>0.881</td>
<td>0.884</td>
<td>0.877</td>
<td>0.884</td>
<td>0.883</td>
<td>0.883</td>
<td>0.883</td>
<td>0.883</td>
</tr>
<tr>
<td>average</td>
<td>0.628</td>
<td>0.657</td>
<td>0.667</td>
<td>0.667</td>
<td>0.614</td>
<td>0.629</td>
<td>0.664</td>
<td>0.637</td>
</tr>
</tbody>
</table>
Fig. 1. Results of the four aggregation functions in terms of (Macro) F1-measure and its associated standard deviation. In the left figure, synthetic datasets are used (m = 5) and, in the right figure, real datasets (Tab. 1 in the main paper). 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 increased.

Fig. 2. Results of the four aggregation functions in terms of (Macro) AUC and its associated standard deviation. In the left figure, synthetic datasets are used (m = 5) and, in the right figure, real datasets (Tab. 1 in the main paper). 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 increased.

### TABLE 3
Results in terms of macroAUC of the four aggregation functions on real crowd datasets, alone and in combination with weighted voting.

<table>
<thead>
<tr>
<th>Dataset</th>
<th>MV</th>
<th>MD</th>
<th>MrD</th>
<th>k-means</th>
<th>Weighted voting + agg</th>
<th>MV</th>
<th>MD</th>
<th>MrD</th>
<th>k-means</th>
</tr>
</thead>
<tbody>
<tr>
<td>adult2</td>
<td>0.767</td>
<td>0.823</td>
<td>0.811</td>
<td>0.804</td>
<td>0.761</td>
<td>0.777</td>
<td>0.785</td>
<td>0.774</td>
<td></td>
</tr>
<tr>
<td>dogs</td>
<td>0.89</td>
<td>0.896</td>
<td>0.897</td>
<td>0.883</td>
<td>0.897</td>
<td>0.899</td>
<td>0.9</td>
<td>0.889</td>
<td></td>
</tr>
<tr>
<td>fej2013</td>
<td>0.811</td>
<td>0.788</td>
<td>0.795</td>
<td>0.814</td>
<td>0.811</td>
<td>0.81</td>
<td>0.81</td>
<td>0.814</td>
<td></td>
</tr>
<tr>
<td>music_genre</td>
<td>0.839</td>
<td>0.842</td>
<td>0.839</td>
<td>0.846</td>
<td>0.887</td>
<td>0.981</td>
<td>0.88</td>
<td>0.89</td>
<td></td>
</tr>
<tr>
<td>sal2013</td>
<td>0.881</td>
<td>0.883</td>
<td>0.888</td>
<td>0.881</td>
<td>0.896</td>
<td>0.883</td>
<td>0.901</td>
<td>0.879</td>
<td></td>
</tr>
<tr>
<td>trvc2010</td>
<td>0.648</td>
<td>0.655</td>
<td>0.655</td>
<td>0.645</td>
<td>0.646</td>
<td>0.655</td>
<td>0.655</td>
<td>0.645</td>
<td></td>
</tr>
<tr>
<td>valence5</td>
<td>0.64</td>
<td>0.709</td>
<td>0.723</td>
<td>0.698</td>
<td>0.615</td>
<td>0.645</td>
<td>0.7</td>
<td>0.649</td>
<td></td>
</tr>
<tr>
<td>weather_sent</td>
<td>0.928</td>
<td>0.931</td>
<td>0.926</td>
<td>0.93</td>
<td>0.929</td>
<td>0.93</td>
<td>0.93</td>
<td>0.929</td>
<td></td>
</tr>
<tr>
<td>wordsim5</td>
<td>0.692</td>
<td>0.773</td>
<td>0.814</td>
<td>0.742</td>
<td>0.662</td>
<td>0.699</td>
<td>0.745</td>
<td>0.682</td>
<td></td>
</tr>
<tr>
<td>average</td>
<td>0.788</td>
<td>0.811</td>
<td>0.817</td>
<td>0.805</td>
<td>0.789</td>
<td>0.798</td>
<td>0.812</td>
<td>0.795</td>
<td></td>
</tr>
</tbody>
</table>
Fig. 3. Proportional difference in terms of (Macro) F1-measure 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 the main paper). 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 the performance difference as the rate of biased annotators \(\gamma\) is increased: 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.

Fig. 4. Proportional difference in terms of (Macro) AUC 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 the main paper). 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 the performance difference as the rate of biased annotators \(\gamma\) is increased: 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.
Fig. 5. Results of the four aggregation functions in terms of a-mean and its associated standard deviation in real datasets (see Tab. 1 in the main paper). Each subfigure considers experiments with different number of annotators, \( t = \{6, 12, 18\} \). Plots are displayed by column depending on the degree of bias, \( \alpha = \{0.0, 0.25, 0.5, 0.75, 1.0\} \), and by row, depending on the relevance, \( \{5, 7\} \), of the real label. Each plot shows performance as the mean imbalance degree (\( \text{ID}_{\text{HE}} \)) increases: moving average of size 3 among the real datasets ordered by \( \text{ID}_{\text{HE}} \). Each plot in these figures expands, from the point of view of class imbalance, the information averaged in a single point in Figure 3 of the main paper.
Fig. 6. Results of the four aggregation functions in terms of macroF1 and its associated standard deviation in real datasets (see Tab. 1 in the main paper). Each subfigure considers experiments with different number of annotators, $t = \{6, 12, 18\}$. Plots are displayed by column depending on the degree of bias, $\alpha = \{0.0, 0.25, 0.5, 0.75, 1.0\}$, and by row, depending on the relevance, $\{5, 7\}$, of the real label. Each plot shows performance as the mean imbalance degree ($ID_{HE}$ [2]) increases, moving average of size 3 among the real datasets ordered by $ID_{HE}$. Each plot in these figures expands, from the point of view of class imbalance, the information averaged in a single point in Figure 1.
Fig. 7. Results of the four aggregation functions in terms of macroAUC and its associated standard deviation in real datasets (see Tab. 1 in the main paper). Each subfigure considers experiments with different number of annotators, $t = \{6, 12, 18\}$. Plots are displayed by column depending on degree of bias, $\alpha = \{0.0, 0.25, 0.5, 0.75, 1.0\}$, and by row, depending on the relevance, $\{5, 7\}$, of the real label. Each plot shows performance as the mean imbalance degree (ID$_{HE}$ [2]) increases: moving average of size 3 among the real datasets ordered by ID$_{HE}$. Each plot in these figures expands, from the point of view of class imbalance, the information averaged in a single point in Figure 2.
Fig. 8. Proportional difference (wMV+agg)/agg of the weighted voting with the four aggregation functions regarding their use alone (in terms of a-mean with real datasets from Tab. 1 in the main paper). In each subfigure, a different number of annotators, $t = \{6, 12, 18\}$, is used. Plots are displayed by column depending on the rate of biased annotators, $\gamma = \{0.0, 0.25, 0.5, 0.75, 1.0\}$, and by row, depending on the relevance, $\{5, 7\}$, of the real label. Each plot shows performance as the mean imbalance degree ($\text{ID}_{\text{HE}}$) increases: moving average of size 3 among the real datasets ordered by $\text{ID}_{\text{HE}}$. A value larger than 1 in the y-axis depicts a scenario where (wMV+agg) outperforms agg alone. Each plot in these figures expands, from the point of view of class imbalance, the information averaged in a single point in Figure 4 of the main paper.
Fig. 9. Proportional difference \((wMV+agg)/agg\) of the weighted voting with the four aggregation functions regarding their use alone (in terms of macroF1 with real datasets from Tab. 1 in the main paper). In each subfigure, a different number of annotators, \(t = \{6, 12, 18\}\), is used. Plots are displayed by column depending on the rate of biased annotators, \(\gamma = \{0.0, 0.25, 0.5, 0.75, 1.0\}\), and by row, depending on the relevance, \(\{5, 7\}\), of the real label. Each plot shows performance as the mean imbalance degree (\(\text{ID}_{\text{HE}}\), \([7]\)) increases: moving average of size 3 among the real datasets ordered by \(\text{ID}_{\text{HE}}\). A value larger than 1 in the y-axis depicts a scenario where \((wMV+agg)\) outperforms \(agg\) alone. Each plot in these figures expands, from the point of view of class imbalance, the information averaged in a single point in Figure 3.
Fig. 10. Proportional difference \( \frac{(wMV+agg)/agg}{agg} \) of the weighted voting with the four aggregation functions regarding their use alone (in terms of macroAUC with real datasets from Tab. 1 in the main paper). In each subfigure, a different number of annotators, \( t = \{6, 12, 18\} \), is used. Plots are displayed by column depending on the rate of biased annotators, \( \gamma = \{0.0, 0.25, 0.5, 0.75, 1.0\} \), and by row, depending on the relevance, \( \{5, 7\} \), of the real label. Each plot shows performance as the mean imbalance degree (ID\text{HE}, \[?\]) increases: moving average of size 3 among the real datasets ordered by ID\text{HE}. A value larger than 1 in the y-axis depicts a scenario where \( (wMV+agg) \) outperforms \( agg \) alone. Each plot in these figures expands, from the point of view of class imbalance, the information averaged in a single point in Figure 4.