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

Jerónimo Hernández-González, Iñaki 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($\{\boldsymbol{a}_1, \boldsymbol{a}_2, \ldots, \boldsymbol{a}_n\}$)
  $\hat{\boldsymbol{h}} \leftarrow new\ tuple(\texttt{nElements:}n)$
  for $j \in \{1, \ldots, n\}$ **do**
    $\boldsymbol{q} \leftarrow new\ tuple(\texttt{nElements:}|\mathcal{C}|)$
    for $c \in \{1, \ldots, |\mathcal{C}|\}$ **do**
      $q_c \leftarrow countsOfLabel(\boldsymbol{a}_j, c)$ ▷ No. annotators providing label c in $\boldsymbol{a}_j$
    end for
    $\boldsymbol{mv} \leftarrow which(\{q_c = \max(\boldsymbol{q})\}_{c=1}^{|\mathcal{C}|})$ ▷ Label(s) which have received the largest number of votes
    if $|\boldsymbol{mv}| = 1$ **then**
      $\hat{h}_j \leftarrow mv_1$ ▷ Each example is assigned to the label c with the largest number of votes (mv_1)
    else if $|\boldsymbol{mv}| > 1$ **then**
      $\hat{h}_j \leftarrow randomSelection(\boldsymbol{mv})$ ▷ Ties are solved randomly: any label with the maximum number of votes
    end if
  end for
  return $\{\hat{h}_1, \hat{h}_2, \ldots, \hat{h}_n\}$
end procedure

Algorithm 2 Pseudocode of the maximum distance (MD) approach.

```
procedure MD({a_1, a_2, ..., a_n})
    Q ← new matrix( nRow:n , nCol:|C| )          ▷ 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)                      ▷ No. annotators providing label c in a_j
        end for
    end for
    q̄ ← meanByRow(Q)                                           ▷ Mean counts (tuple)
    ĥ ← new tuple(nElements:n)
    for j ∈ {1, ..., n} do
        mv ← which({q_c − q̂_c = max(q − q̂)}_{c=1}^{|C|})        ▷ Label(s) which have received the largest number of votes
                                                               ▷ in comparison with its mean
        if |mv| = 1 then
            ĥ_j ← mv_1          ▷ Each example is assigned to the label c with the largest number of votes (mv_1)
        else if |mv| > 1 then
            ĥ_j ← randomSelection(mv)     ▷ Ties are solved randomly: any label with the maximum number of votes
        end if
    end for
    return {ĥ_1, ĥ_2, ..., ĥ_n}
end procedure
```

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

```
procedure MrD({a_1, a_2, ..., a_n})
    Q ← new matrix( nRow:n , nCol:|C| )          ▷ 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_ic ← countsOfLabel(a_j, c)                      ▷ No. annotators providing label c in a_j
        end for
    end for
    q̄ ← meanByRow(Q)                                           ▷ Mean counts (tuple)
    ĥ ← new tuple(nElements:n)
    for j ∈ {1, ..., n} do
        mv ← which({q_c/q̂_c = max(q/q̂)}_{c=1}^{|C|})            ▷ Label(s) which have received the largest number of votes
                                                               ▷ relative to its mean
        if |mv| = 1 then
            ĥ_j ← mv_1          ▷ Each example is assigned to the label c with the largest number of votes (mv_1)
        else if |mv| > 1 then
            ĥ_j ← randomSelection(mv)     ▷ Ties are solved randomly: any label with the maximum number of votes
        end if
    end for
    return {ĥ_1, ĥ_2, ..., ĥ_n}
end procedure
```

Algorithm 4 Pseudocode of the k-means based approach.

```
procedure K-MEANS({a_1, a_2, ..., a_n})
    Q ← new matrix( nRow:n , nCol:|C| + 1 )      ▷ 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)                      ▷ No. annotators providing label c in a_j
        end for
        q_j(|C|+1) ← Σ_{c=2}^{|C|} q_jc − q_j(c−1)
    end for
    iCentroids ← {arg max_{j∈1,...,n} q_jc}_{c=1}^{|C|}       ▷ Each centroid represents a class label c (k = |C|)
    ĥ ← Kmeans(Q, iCentroids, k = |C|)                         ▷ Assign each example to a centroid
    return {ĥ_1, ĥ_2, ..., ĥ_n}     ▷ 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($\{\boldsymbol{a}_1, \boldsymbol{a}_2, \ldots, \boldsymbol{a}_n\}$)
  $W \leftarrow new\ matrix($ `nRow:t, nCol:`$|\mathcal{C}|$ $)$ ▷ w_c^l is the cell in the intersection of the l-th row and the c-th column of W
  $\hat{\boldsymbol{h}} = agg(\{\boldsymbol{a}_1, \boldsymbol{a}_2, \ldots, \boldsymbol{a}_n\})$
  for $c \in \{1, \ldots, |\mathcal{C}|\}$ **do**
    $\hat{\boldsymbol{h}}^c \leftarrow examplesOfLabel(\hat{\boldsymbol{h}}, c)$ ▷ Examples which have been assigned by agg to label c
    for $l \in \{1, \ldots, t\}$ **do**
      $\boldsymbol{a}_l^c \leftarrow annotationsOfLabel(\{\boldsymbol{a}_1, \boldsymbol{a}_2, \ldots, \boldsymbol{a}_n\}, l, c)$ ▷ Examples assigned by labeler l to label c
      $w_c^l \leftarrow |\hat{\boldsymbol{h}}^c \cap \boldsymbol{a}_l^c| / |\hat{\boldsymbol{h}}^c|$
    end for
  end for
  $\hat{\boldsymbol{h}} \leftarrow new\ tuple($`nElements:`$n)$
  for $j \in \{1, \ldots, n\}$ **do**
    $\boldsymbol{q} \leftarrow new\ tuple($`nElements:`$|\mathcal{C}|)$
    for $c \in \{1, \ldots, |\mathcal{C}|\}$ **do**
      $q_c \leftarrow \sum_{l=1}^{t} w_c^l \cdot \mathbb{I}[a_l^j = c]$ ▷ Summation of the weight of annotators which provided label c
    end for
    $\boldsymbol{mv} \leftarrow which(\{q_c = \max(\boldsymbol{q})\}_{c=1}^{|\mathcal{C}|})$ ▷ Label(s) which have received the largest number of votes
    if $|\boldsymbol{mv}| = 1$ **then**
      $\hat{h}_j \leftarrow mv_1$ ▷ Each example is assigned to the label c with the largest number of votes (mv_1)
    else if $|\boldsymbol{mv}| > 1$ **then**
      $\hat{h}_j \leftarrow randomSelection(\boldsymbol{mv})$ ▷ Ties are solved randomly: any label with the maximum number of votes
    end if
  end for
  return $\{\hat{h}_1, \hat{h}_2, \ldots, \hat{h}_n\}$
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 $\bar{\boldsymbol{q}} = \{0.54, 0.31, 0.15\}$. In case of a tie, the first class label is selected.

Annotations						Proportions of annots.			$\boldsymbol{q} - \bar{\boldsymbol{q}}$			$\boldsymbol{q}/\bar{\boldsymbol{q}}$			Results		
L_1	L_2	L_3	L_4	L_5	L_6	q_1	q_2	q_3	$c=1$	$c=2$	$c=3$	$c=1$	$c=2$	$c=3$	MV	MD	MrD
1	1	3	1	2	1	0.667	0.167	0.167	0.127	-0.143	0.017	1.235	0.539	1.113	1	1	1
1	1	2	2	1	2	0.500	0.500	0.000	0.040	0.190	-0.150	0.926	1.613	0.000	1	2	2
2	3	2	3	2	1	0.167	0.500	0.333	-0.373	0.190	0.183	0.309	1.613	2.220	2	2	3
3	2	1	2	2	1	0.333	0.500	0.167	-0.207	0.190	0.017	0.617	1.613	1.113	2	2	2
1	2	1	1	1	2	0.667	0.333	0.000	0.127	0.023	-0.150	1.235	1.074	0.000	1	1	1
3	2	3	3	2	2	0.000	0.500	0.500	-0.540	0.190	0.350	0.000	1.613	3.333	2	3	3
3	1	2	1	1	2	0.500	0.333	0.167	-0.040	0.023	0.017	0.926	1.074	1.113	1	2	3
1	1	1	2	1	1	0.833	0.167	0.000	0.290	-0.143	-0.150	1.543	0.539	0.000	1	1	1
3	1	1	1	3	1	0.667	0.000	0.333	0.127	-0.310	0.183	1.235	0.000	2.220	1	3	3
3	3	1	2	1	3	0.333	0.167	0.500	-0.207	-0.143	0.350	0.617	0.539	3.333	3	3	3
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
3	1	1	2	1	3	0.500	0.167	0.333	-0.040	-0.143	0.183	0.926	0.539	2.220	1	3	3
1	2	1	2	2	3	0.333	0.500	0.167	-0.207	0.190	0.017	0.617	1.613	1.113	2	2	2

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

					Weighted voting + agg			
Dataset	MV	MD	MrD	k-means	MV	MD	MrD	k-means
adult2	0.636	0.706	0.674	0.688	0.627	0.652	0.667	0.652
dogs	0.823	0.831	0.833	0.813	0.836	0.839	0.841	0.823
fej2013	0.507	0.477	0.559	0.543	0.508	0.507	0.507	0.515
music_genre	0.713	0.709	0.703	0.722	0.812	0.789	0.785	0.815
saj2013	0.785	0.78	0.795	0.793	0.808	0.78	0.806	0.777
trec2010	0.461	0.468	0.469	0.459	0.462	0.469	0.469	0.463
valence5	0.416	0.491	0.514	0.501	0.221	0.375	0.495	0.426
weather_sent	0.881	0.884	0.877	0.884	0.883	0.883	0.883	0.883
wordsim5	0.429	0.57	0.617	0.602	0.369	0.368	0.522	0.38
average	0.628	0.657	0.671	0.667	0.614	0.629	0.664	0.637

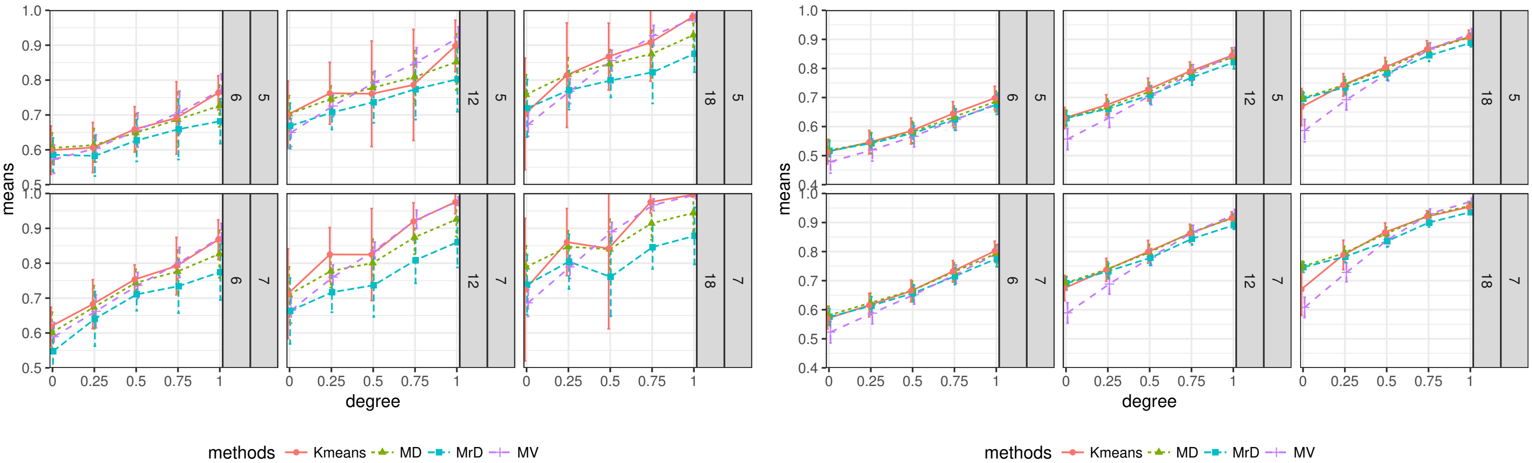


Fig. 1. Results of the four aggregation functions in terms of (Macro) F1-measure and its associated standard deviation. In the left figure, synthetic datasets are used ($m = 5$) and, in the right figure, real datasets (Tab. 1 in the main paper). In both figures, plots are displayed by column depending 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.

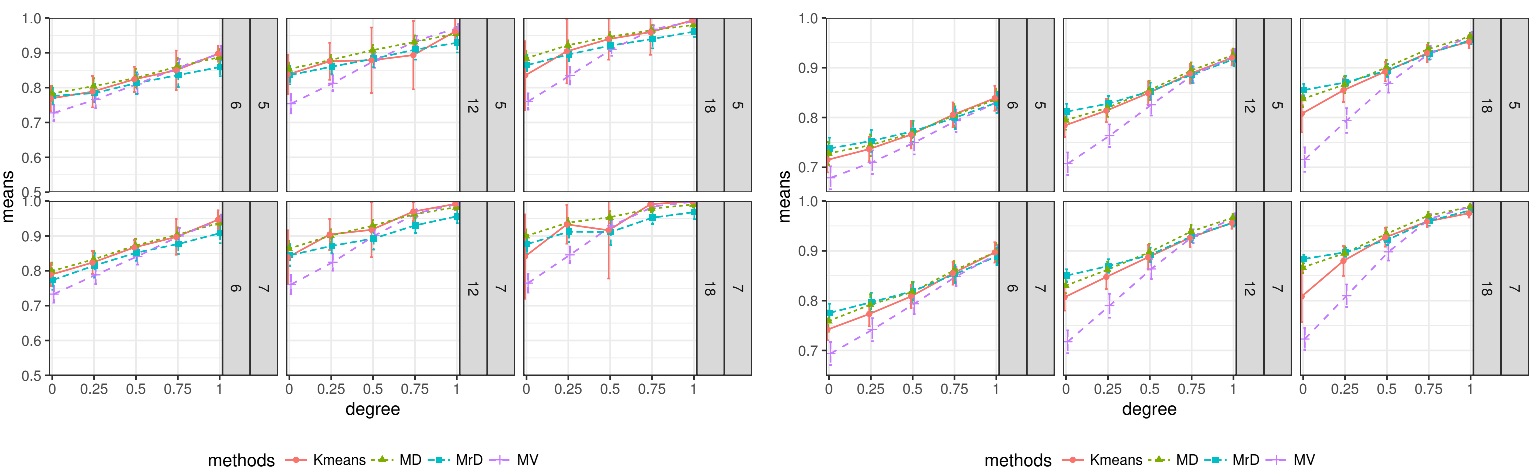


Fig. 2. Results of the four aggregation functions in terms of (Macro) AUC and its associated standard deviation. In the left figure, synthetic datasets are used ($m = 5$) and, in the right figure, real datasets (Tab. 1 in the main paper). In both figures, plots are displayed by column depending on the 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.

					Weighted voting + *agg*			
Dataset	MV	MD	MrD	k-means	MV	MD	MrD	k-means
adult2	0.767	0.823	0.811	0.804	0.761	0.777	0.785	0.774
dogs	0.89	0.896	0.897	0.883	0.897	0.899	0.9	0.889
fej2013	0.811	0.788	0.795	0.814	0.811	0.81	0.81	0.814
music_genre	0.839	0.842	0.839	0.846	0.887	0.881	0.88	0.89
saj2013	0.881	0.883	0.888	0.881	0.896	0.883	0.901	0.879
trec2010	0.648	0.655	0.655	0.645	0.646	0.655	0.655	0.645
valence5	0.64	0.709	0.723	0.698	0.615	0.645	0.7	0.649
weather_sent	0.928	0.931	0.926	0.93	0.929	0.93	0.93	0.929
wordsim5	0.692	0.773	0.814	0.742	0.662	0.699	0.745	0.682
average	0.788	0.811	0.817	0.805	0.789	0.798	0.812	0.795

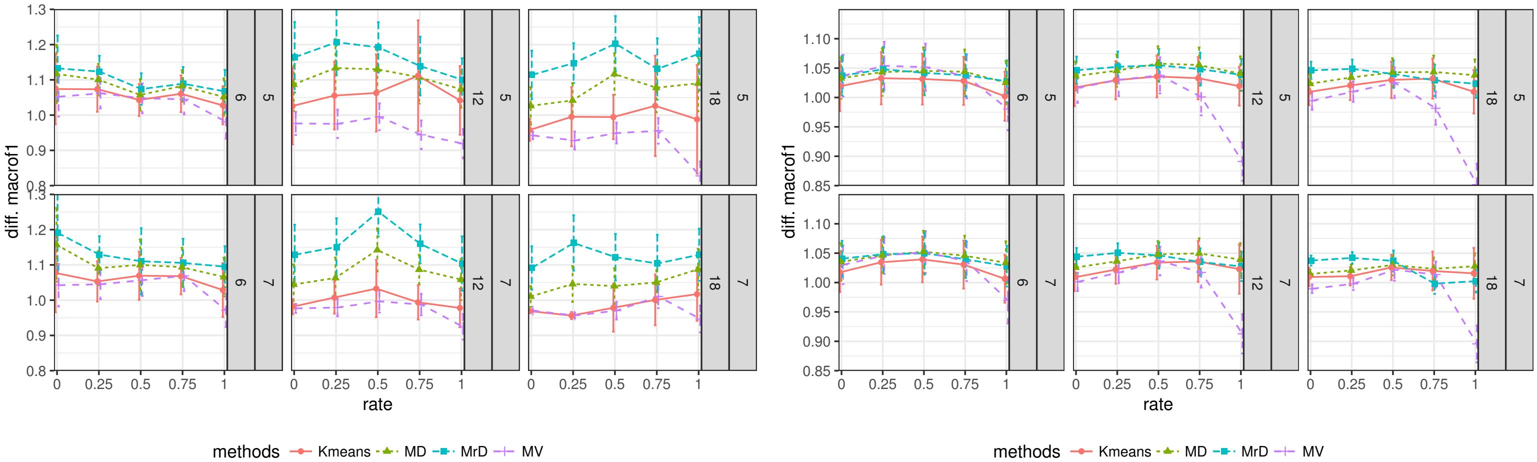


Fig. 3. Proportional difference in terms of (Macro) F1-measure of the results of wMV in combination with the four aggregation functions regarding the use of the four aggregators alone. In the left figure, synthetic datasets are used ($m = 5$) and, in the right figure, real datasets (Tab. 1 in the main 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 (γ) 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.

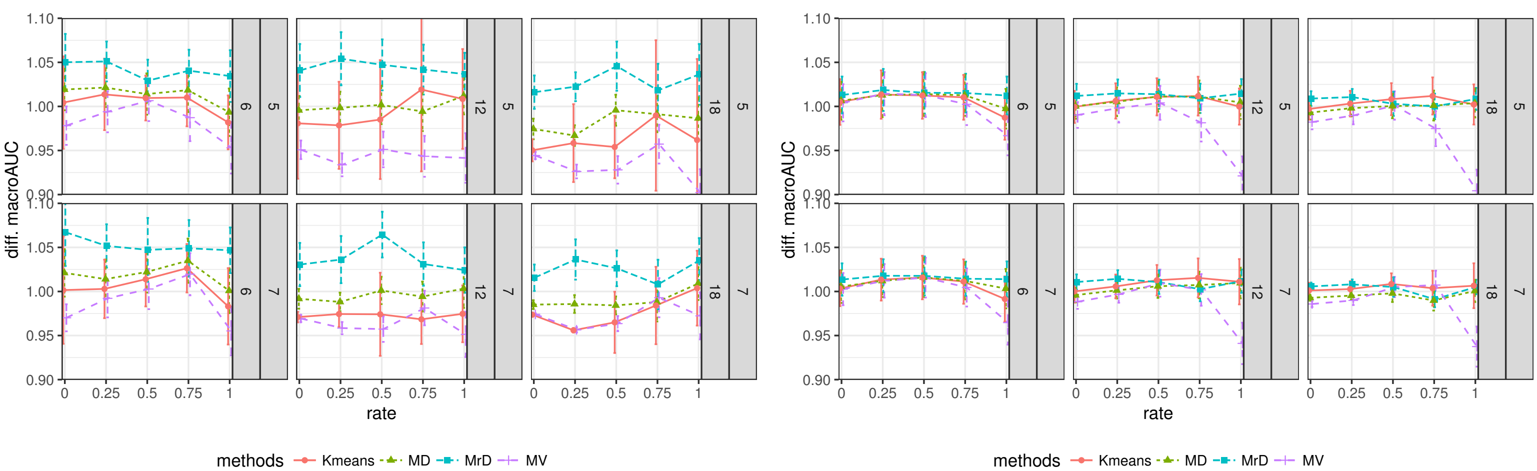
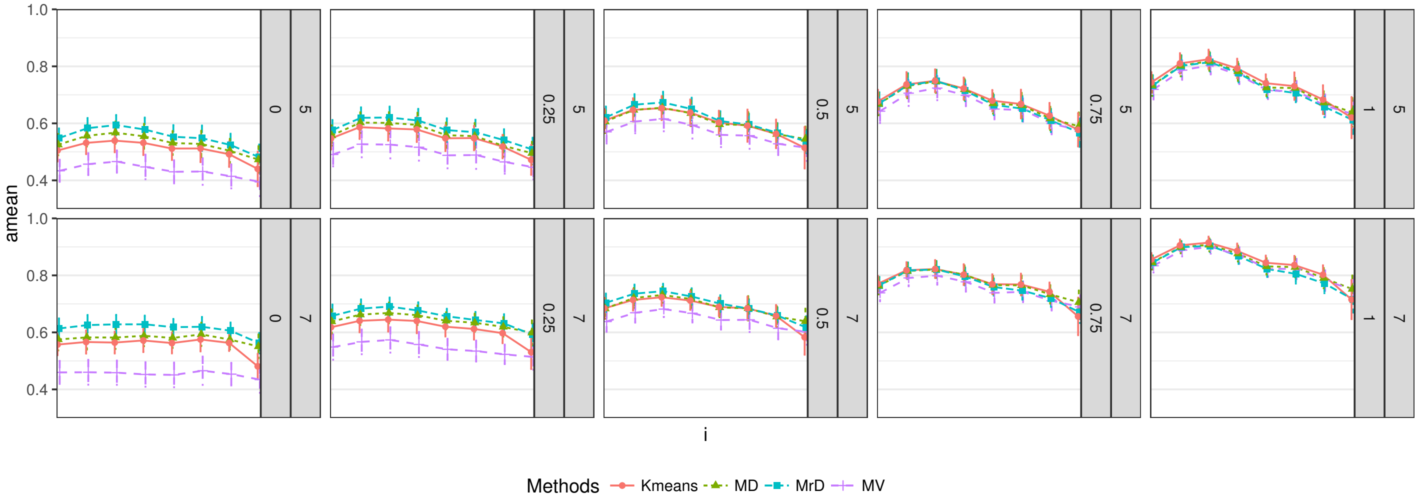
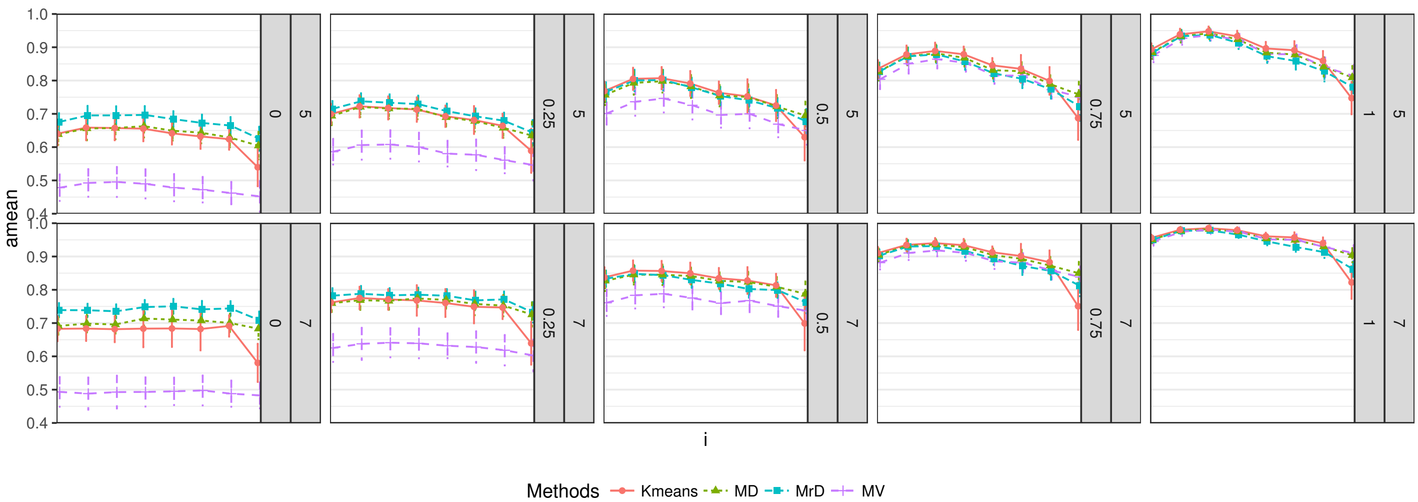


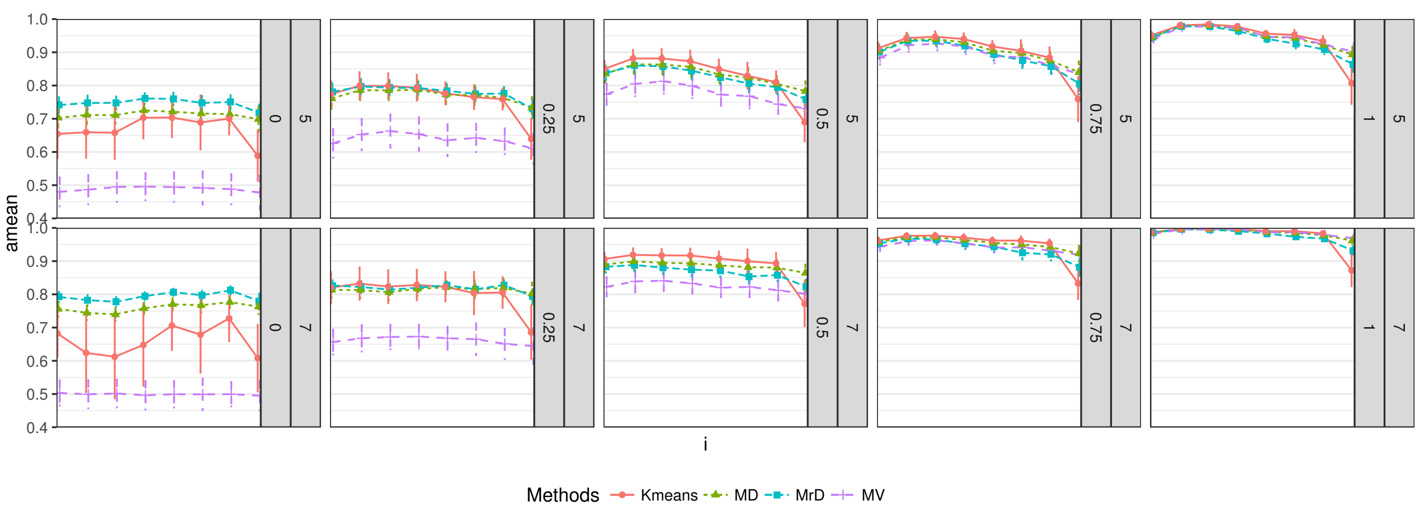
Fig. 4. Proportional difference in terms of (Macro) AUC of the results of wMV in combination with the four aggregation functions regarding the use of the four aggregators alone. In the left figure, synthetic datasets are used ($m = 5$) and, in the right figure, real datasets (Tab. 1 in the main paper). 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 (γ) 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.



(a) $t = 6$

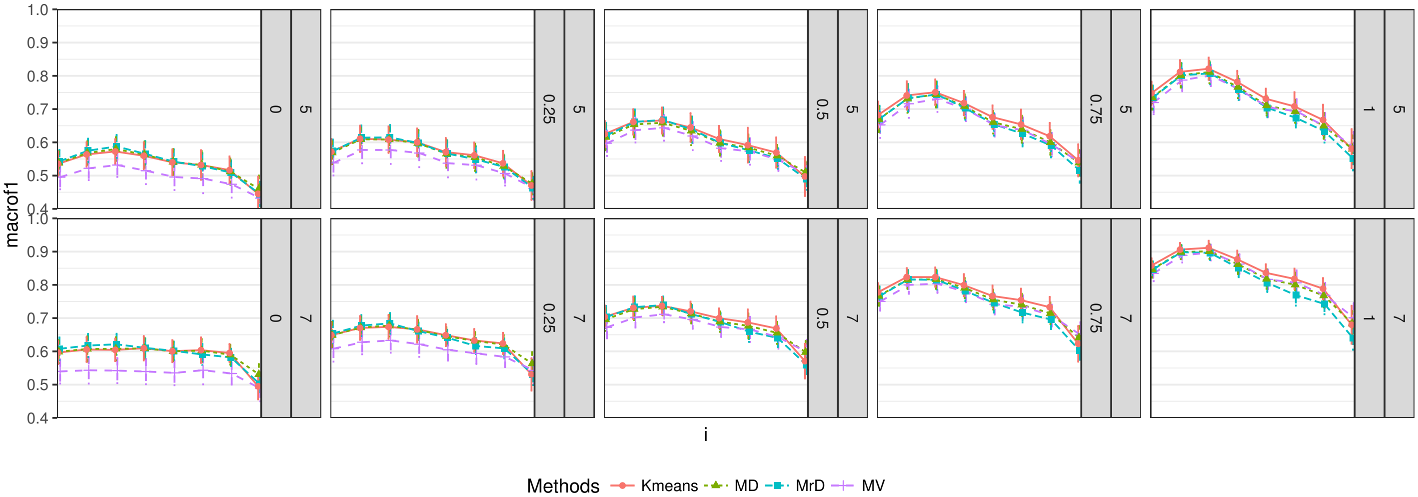


(b) $t = 12$

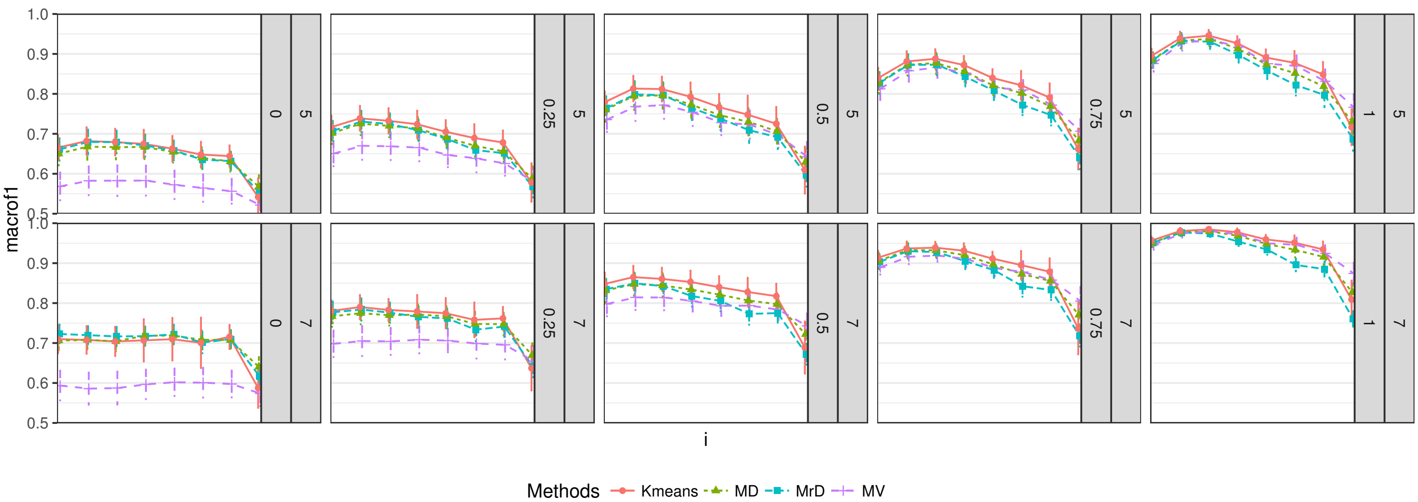


(c) $t = 18$

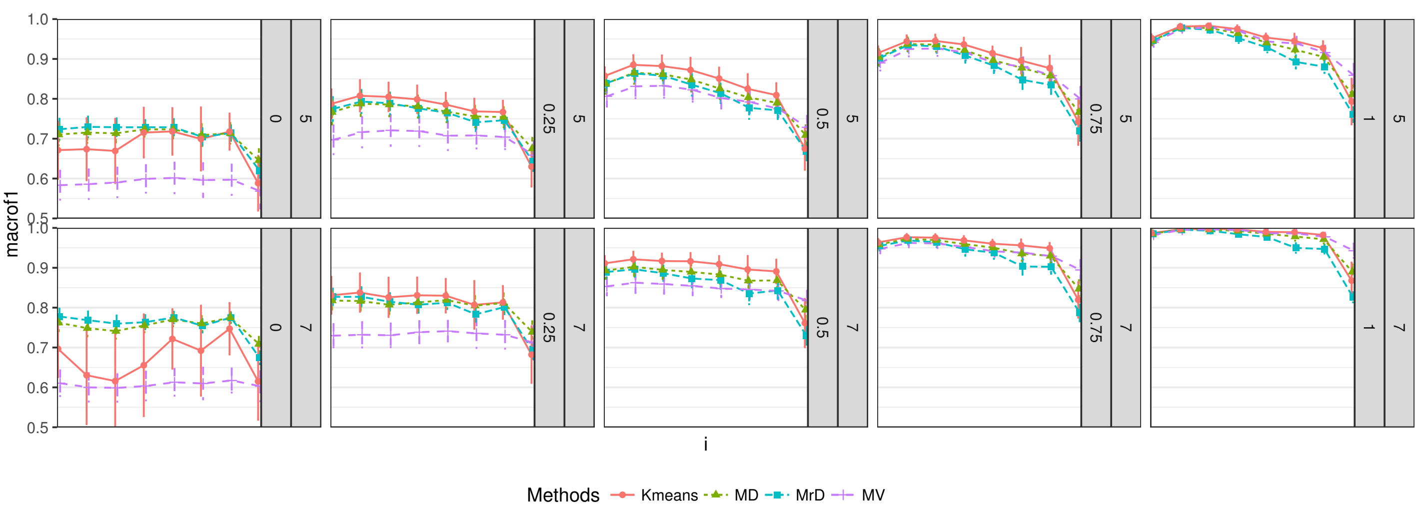
Fig. 5. Results of the four aggregation functions in terms of a-mean and its associated standard deviation in real datasets (see Tab. 1 in the main paper). Each subfigure considers experiments with different number of annotators, $t = \{6, 12, 18\}$. Plots are displayed by column depending on the degree 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 (ID_{HE}, [?]) 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 3 of the main paper.



(a) $t = 6$

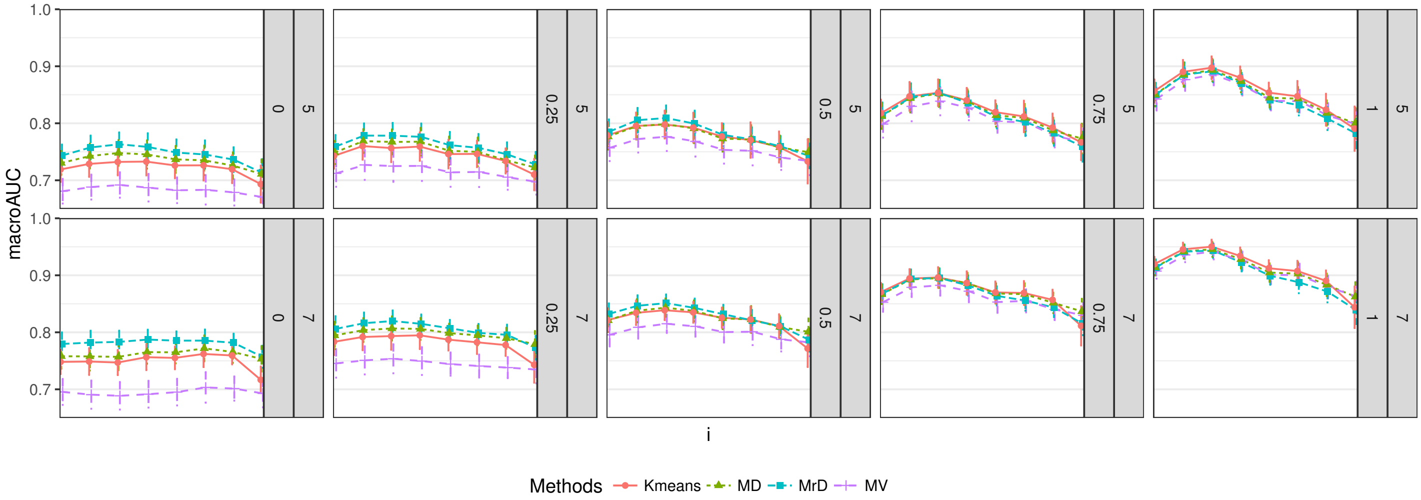


(b) $t = 12$

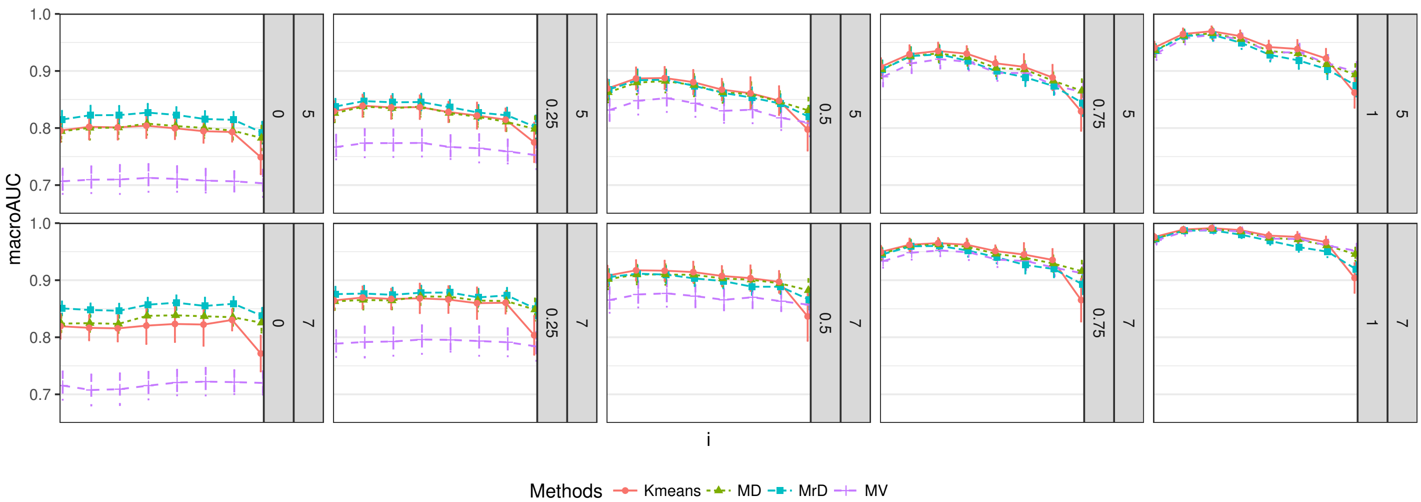


(c) $t = 18$

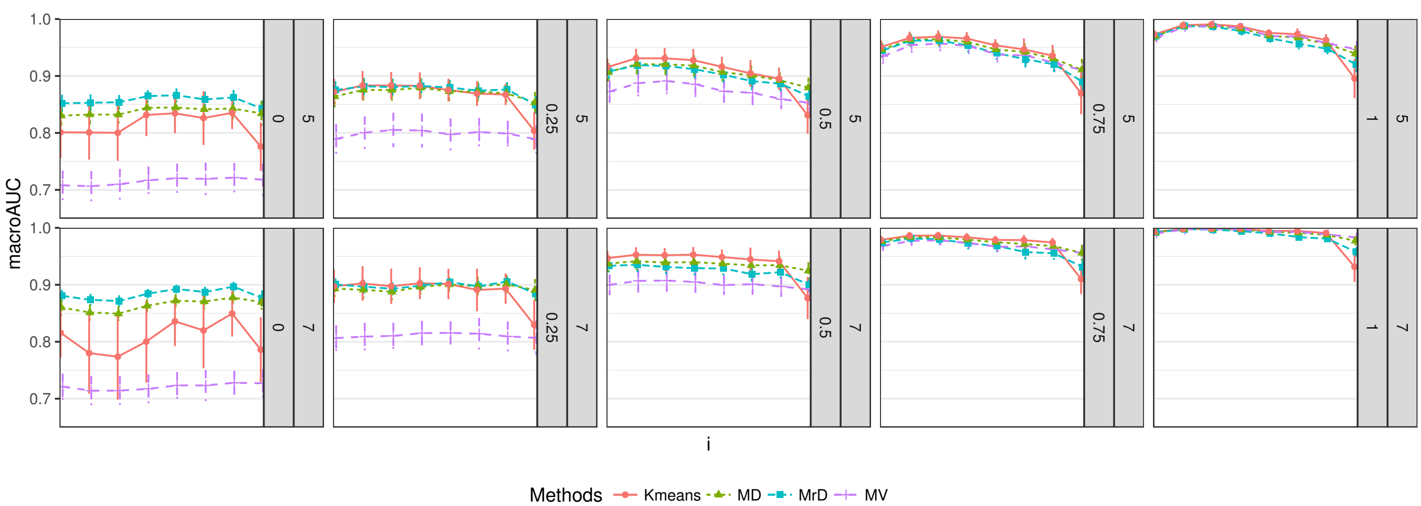
Fig. 6. Results of the four aggregation functions in terms of macroF1 and its associated standard deviation in real datasets (see Tab. 1 in the main paper). Each subfigure considers experiments with different number of annotators, $t = \{6, 12, 18\}$. Plots are displayed by column depending on the degree 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}, [**?**]) 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.



(a) $t = 6$

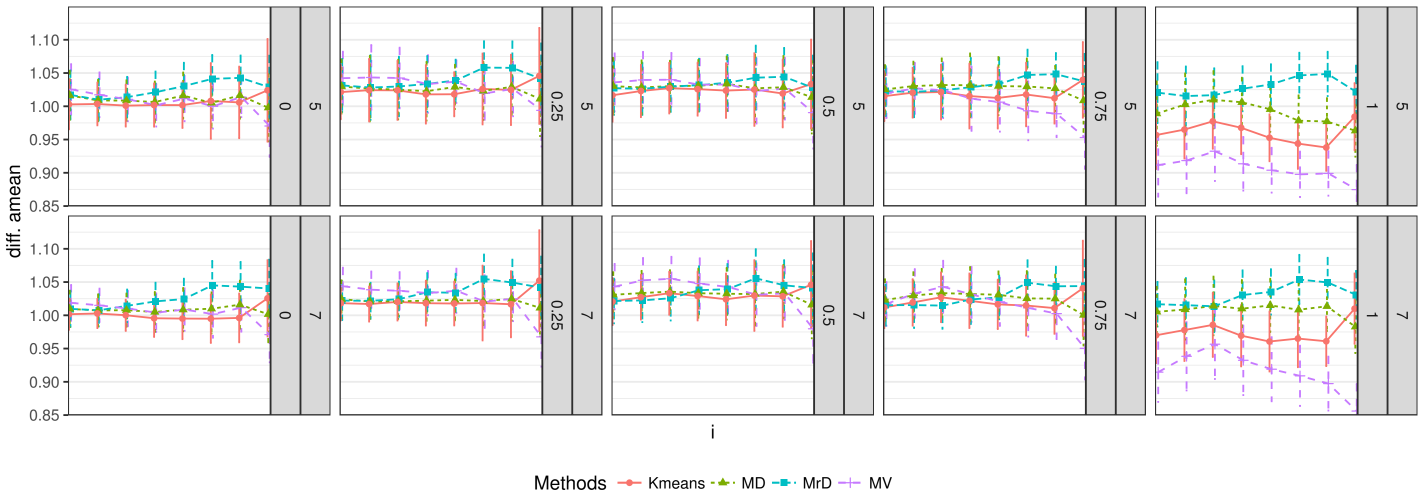


(b) $t = 12$

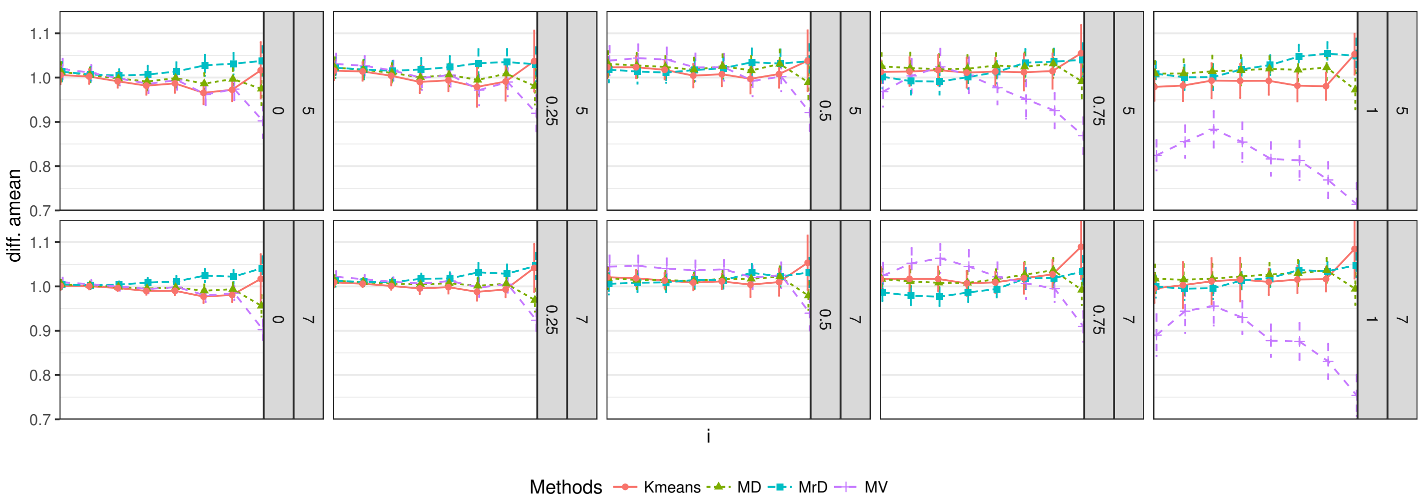


(c) $t = 18$

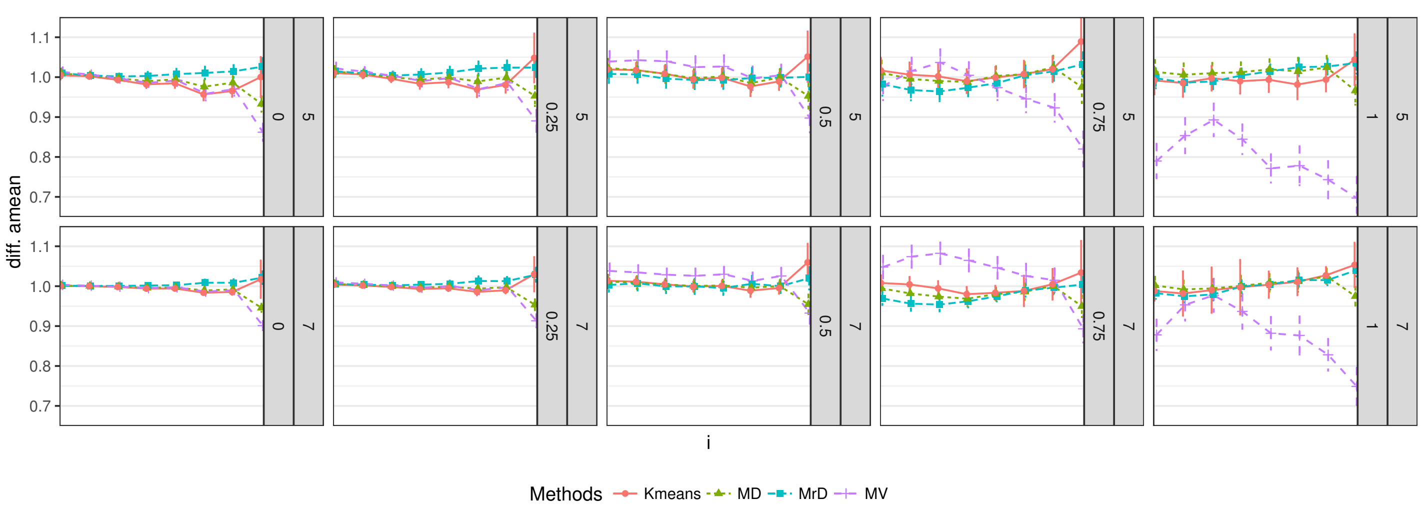
Fig. 7. Results of the four aggregation functions in terms of macroAUC and its associated standard deviation in real datasets (see Tab. 1 in the main paper). Each subfigure considers experiments with different number of annotators, $t = \{6, 12, 18\}$. Plots are displayed by column depending on degree 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}, [?]) 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.



(a) $t = 6$

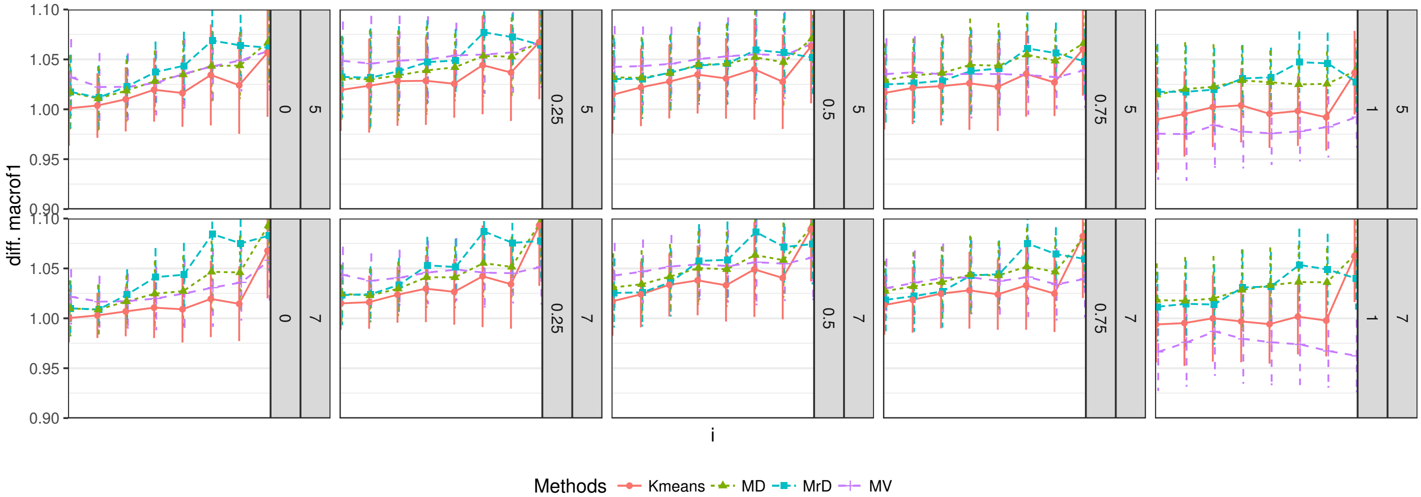


(b) $t = 12$

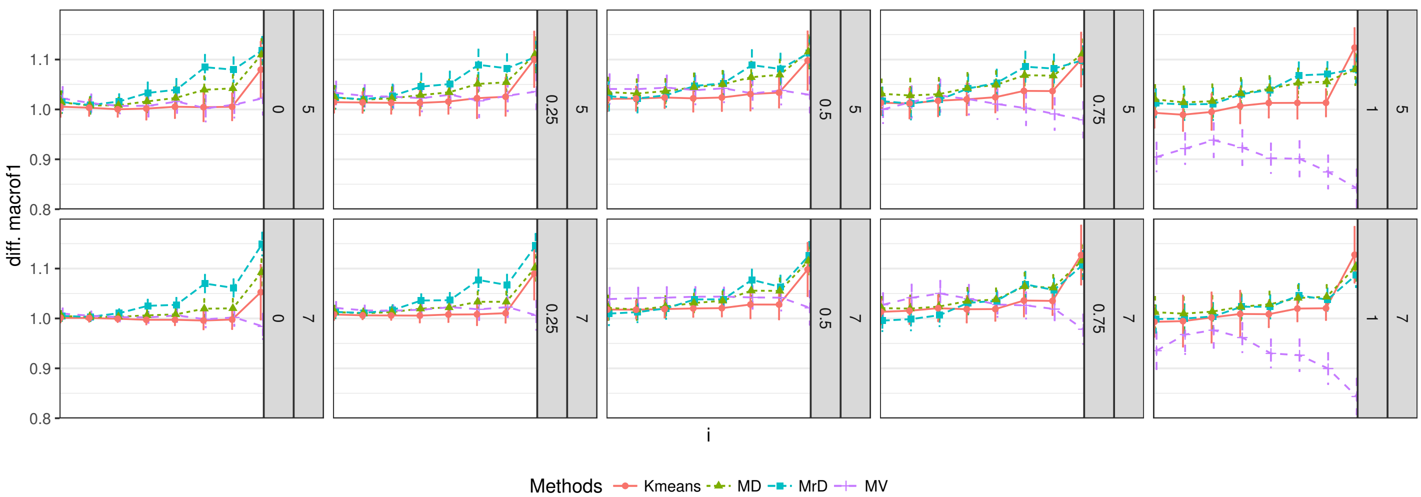


(c) $t = 18$

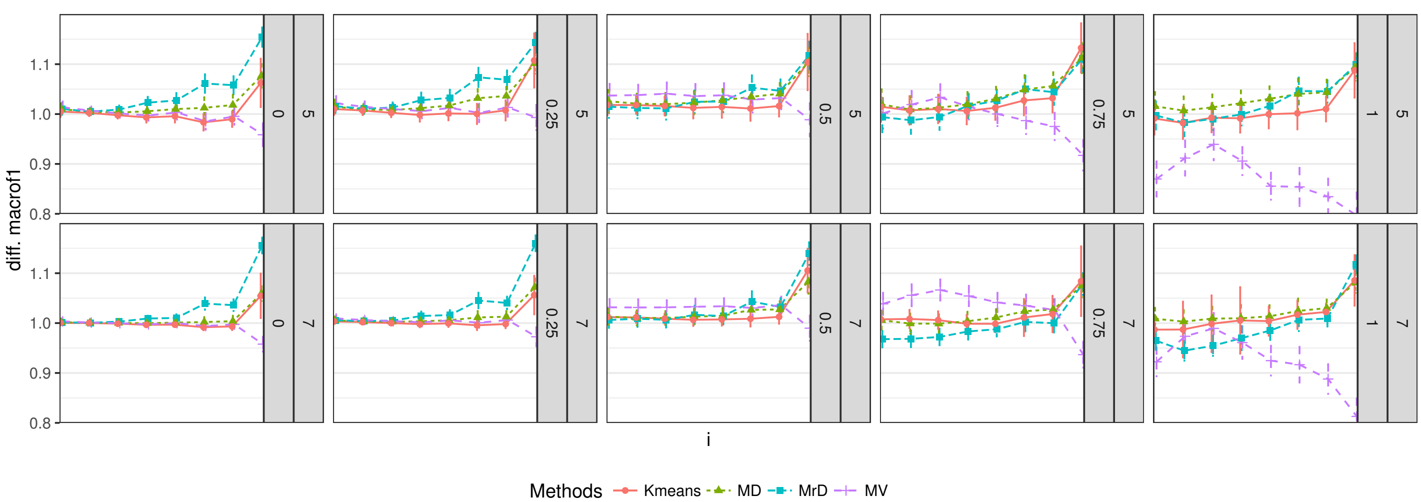
Fig. 8. Proportional difference (wMV+*agg*)/*agg* of the weighted voting with the four aggregation functions regarding their use alone (in terms of a-mean with real datasets from Tab. 1 in the main paper). In each subfigure, a different number of annotators, $t = \{6, 12, 18\}$, is used. Plots are displayed 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 (ID_{HE}, [?]) increases: moving average of size 3 among the real datasets ordered by ID_{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.



(a) $t = 6$

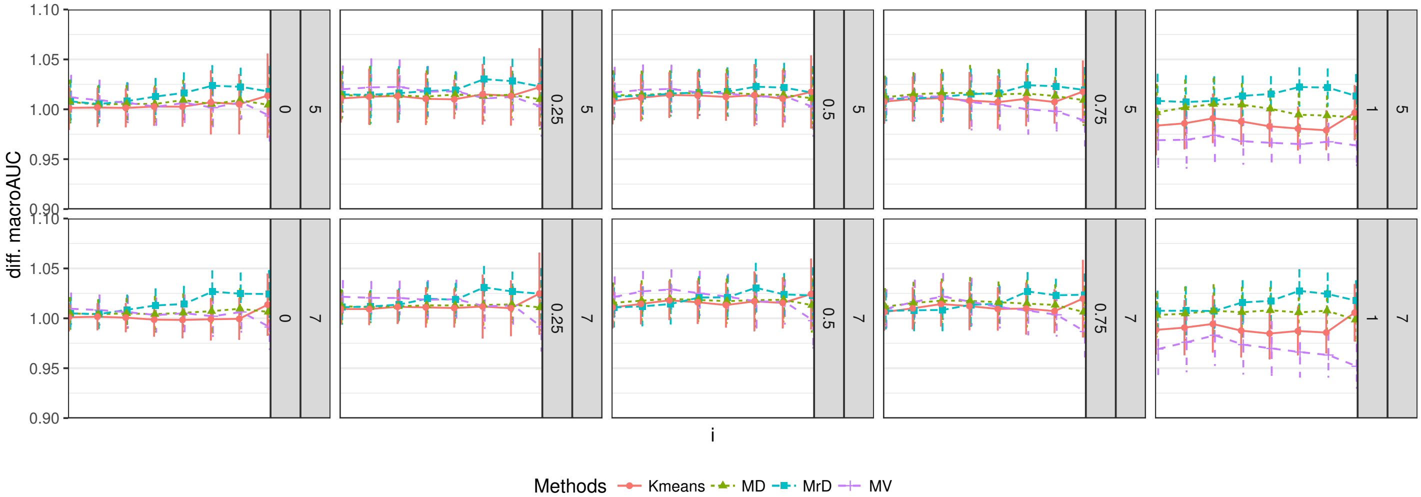


(b) $t = 12$

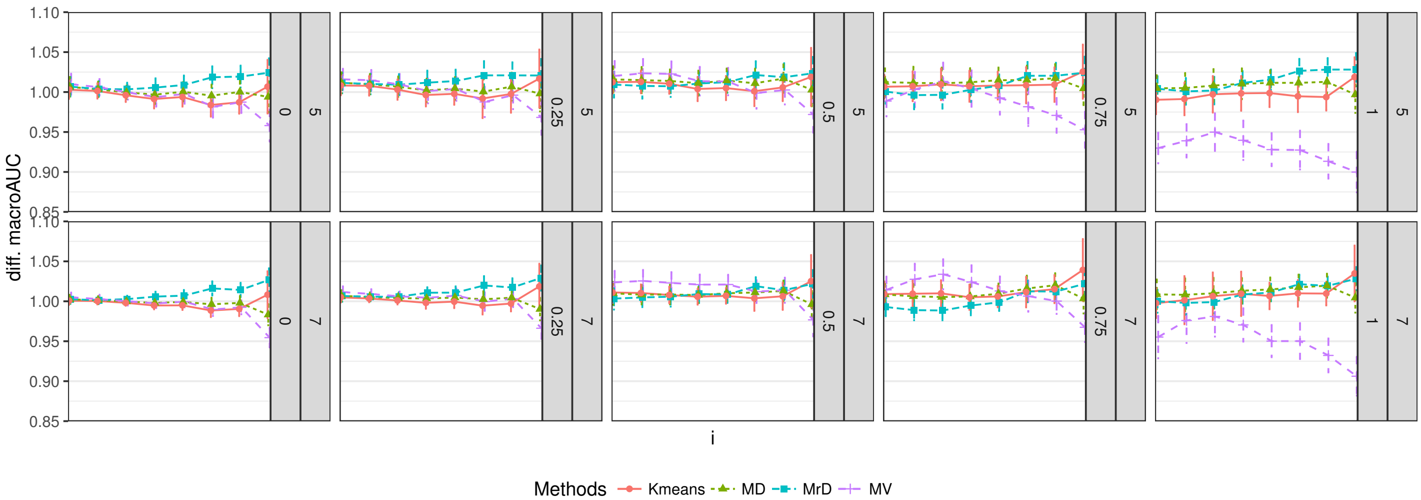


(c) $t = 18$

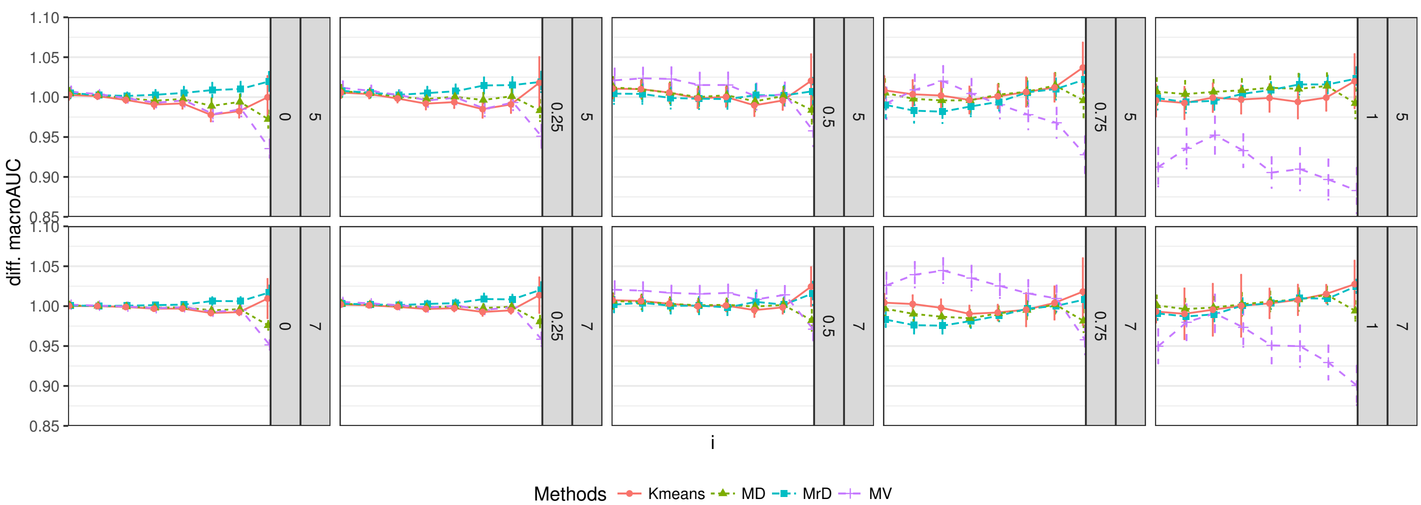
Fig. 9. Proportional difference (wMV+*agg*)/*agg* of the weighted voting with the four aggregation functions regarding their use alone (in terms of macroF1 with real datasets from Tab. 1 in the main paper). In each subfigure, a different number of annotators, $t = \{6, 12, 18\}$, is used. Plots are displayed 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 (ID_{HE}, [?]) increases: moving average of size 3 among the real datasets ordered by ID_{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.



(a) $t = 6$



(b) $t = 12$



(c) $t = 18$

Fig. 10. Proportional difference (wMV+*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_{HE}, [?]) increases: moving average of size 3 among the real datasets ordered by ID_{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.