Supplementary Material for ”Machine learning from crowds using candidate set-based labelling”

Iker Beñaran-Muñoz[1], Jerónimo Hernández-González[2], and Aritz Pérez[3]

[1]Basque Center for Applied Mathematics, Bilbao, Spain
[2]Departament de Matemàtiques i Informàtica, Universitat de Barcelona (UB), Barcelona, Spain
[3]Basque Center for Applied Mathematics, Bilbao, Spain

May 6, 2022

In this document we present the material that could not be included in the original manuscript due to space restrictions of the journal. The different pieces are shown here in the same order as they are mentioned in the paper.

Notation

In Table 1, all the notation used in the paper is described.

Figures with additional datasets

In Figures 1, 2 and 3, results obtained with other datasets that are not shown in the manuscript are displayed. Results with 3 supervised datasets from UCI repository (`http://archive.ics.uci.edu/ml`) are shown: *Arrhythmia* $(452, 13)$, *Pendigits* $(10992, 10)$ and *Satimage* $(6435, 6)$, with numbers meaning no. instances n, and no. classes r.

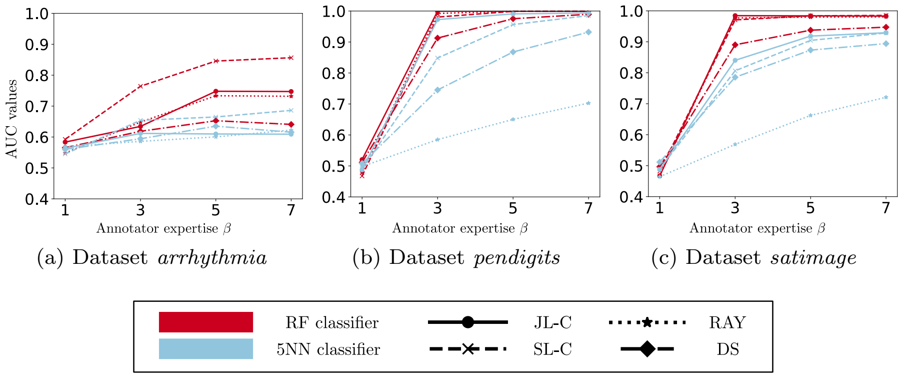
We simulate different numbers of annotators $m \in \{3, 5, 7, 9\}$, and different degrees of expertise for them $\beta \in \{1, 3, 5, 7\}$. For candidate labelling generation, the proportion of sampled labels takes values $prop \in \{0.1, 0.3, 0.5, 0.7\}$. We have used two classifiers from very different families from *sklearn 0.22.1* with default parameters: 5-Nearest Neighbour (5NN) and Random Forest (RF).

The models are evaluated using the area under the ROC curve (AUC). It is estimated using stratified 5-fold cross-validation, where the test sets are fully supervised.

Figure 1 shows the impact of the expertise (β parameter) of the annotators on the performance of the methods, Figure 2 shows the impact of the number of annotators (m) on the performance of the methods, and Figure 3 shows the effect of the maximum candidate set size ($prop$) in the performance.

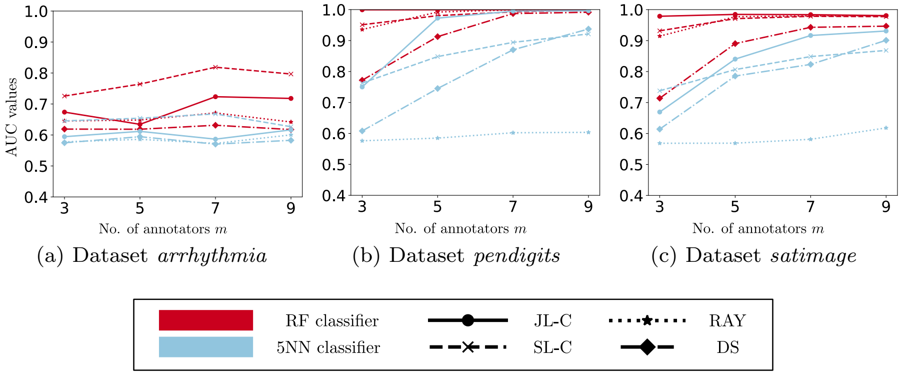
X	Descriptive variable
d	Dimension of the descriptive variable
x	Instance of X
Ω_X	Feature space
C	Class variable
c	Class label
Ω_C	Set of possible class labels
A	Set of available annotators
a	Annotator from A
l_x^a	Label provided by annotator a for instance x (full labelling context)
L_x^a	Candidate set provided by annotator a for instance x (candidate labelling context)
$\mathcal{L}_x$	Labelling for instance x
$\mathcal{L}$	Set of labellings for the whole training set
w_x	Candidate voting estimate
ω	Candidate voting function
α_{ck}^a	(Parameter) Probability that annotator a includes class label k in the candidate set for an instance of true class c
$\boldsymbol{\alpha}$	Set of all α_{ck}^a parameters
$q_{\hat{\boldsymbol{\alpha}}}(c\|x)$	Estimate of the probability that instance x belongs to class c, based on the set of parameters $\boldsymbol{\alpha}$ (in the SL-C method)
h	Probabilistic classifier
θ	Parameter set of the probabilistic classifier
$q_{\hat{\boldsymbol{\alpha}},\hat{\theta}}(c\|x)$	Estimate of the probability that instance x belongs to class c, based on the set of parameters $(\hat{\boldsymbol{\alpha}}, \hat{\theta})$ (in the JL-C method)
n	Total number of instances
r	Number of classes
m	Number of annotators
β	Label generation parameter that represents annotator expertise
$prop$	Proportion of sampled labels over the possible class labels when generating candidate sets
g_a	Probability distribution that represents annotator a in the label generation process
$\mathcal{C}_x$	Set of partial labels associated to instance x

Table 1: Notation used in the work.



(a) Dataset *arrhythmia* (b) Dataset *pendigits* (c) Dataset *satimage*

Figure 1: Experimental results throughout different values of the parameter β (annotator expertise), in terms of AUC metric, within different datasets (subplots). Results with classifiers RF and 5NN are displayed in orange and red colors, respectively. A different line style and marker is used for each method (SL-C, JL-C, RAY , DS). The rest of generative parameters are fixed to $m = 5$ and $prop = 0.5$.

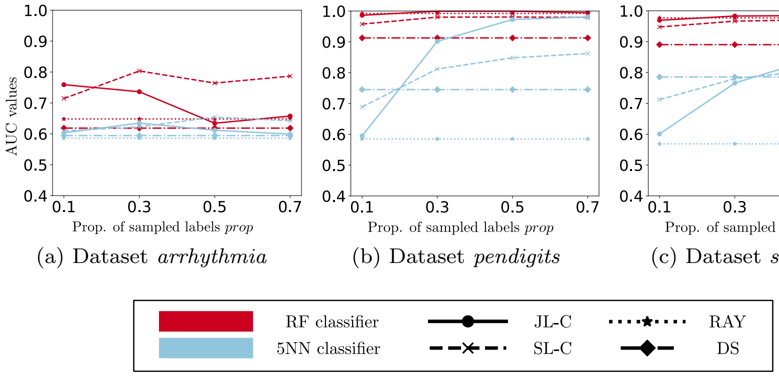


(a) Dataset *arrhythmia* (b) Dataset *pendigits* (c) Dataset *satimage*

Figure 2: Experimental results throughout different values of the parameter m (number of annotators), in terms of AUC metric, within different datasets (subplots). Results with classifiers RF and 5NN are displayed in orange and red colors, respectively. A different line style and marker is used for each method (SL-C, JL-C, RAY , DS). The rest of generative parameters are fixed to $\beta = 3$ and $prop = 0.5$.

Figures with additional configurations

In Figures 1, 2 and 3, results obtained with the fixed parameter value $\beta = 5$ are shown. We use 9 fully labelled datasets from UCI repository (`http://archive.`



(a) Dataset *arrhythmia*   (b) Dataset *pendigits*   (c) Dataset *satimage*

Figure 3: Experimental results throughout different values of the parameter *prop* (flexibility of the annotators), in terms of AUC metric, within different datasets (subplots). Results with classifiers RF and 5NN are displayed in orange and red colours, respectively. A different line style and marker is used for each method (SL-C, JL-C, RAY , DS). The rest of generative parameters are fixed to $\beta = 3$ and $m = 5$.

`ics.uci.edu/ml`): *Arrhythmia* $(452, 13)$, *Dermatology* $(366, 6)$, *Glass* $(214, 6)$, *Pendigits* $(10992, 10)$, *Satimage* $(6435, 6)$, *Segment* $(2310, 7)$, *Svmguide4* $(612, 6)$, *Vehicle* $(846, 4)$, and *Vowel* $(990, 11)$, and 3 partially labelled datasets [1] []: *Birdac* $(3718, 13)$, *Lost* $(1122, 14)$ and *MSRCv2* $(1758, 23)$, with numbers meaning number of instances n, and number of classes r.

We simulate different numbers of annotators $m \in \{3, 5, 7, 9\}$, and different degrees of expertise for them $\beta \in \{1, 3, 5, 7\}$. For candidate labelling generation, the proportion of sampled labels takes values $prop \in \{0.1, 0.3, 0.5, 0.7\}$. We have used two classifiers from very different families from *sklearn 0.22.1* with default parameters: 5-Nearest Neighbour (5NN) and Random Forest (RF).

The models are evaluated using the area under the ROC curve (AUC). It is estimated using stratified 5-fold cross-validation, where the test sets are fully supervised.

Figure 4 shows the impact of the number of annotators (m) on the performance of the methods, and Figure 5 shows the effect of the maximum candidate set size ($prop$) in the performance.

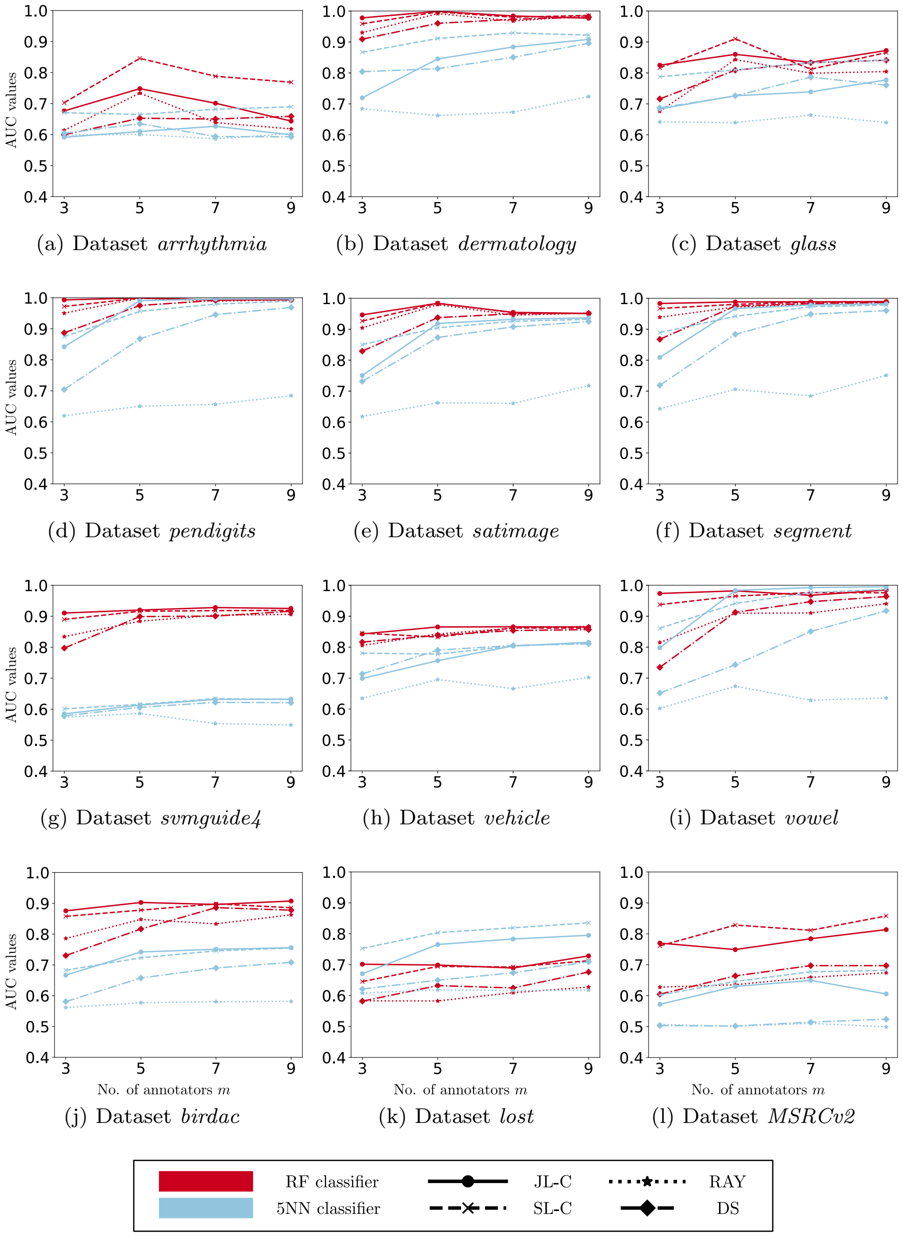
Scalability test

We have performed a scalability test by using subsets of increasing size of the *pendigits* dataset (the largest one among the considered datasets) and applying our methods SL-C and JL-C on them. Each subset has a portion of the original instances, ranging from 0.1 (10% of the instances are preserved) to 1 (the entire dataset is used). Stratified sampling is performed to keep the class proportions. The time taken for a run of the EM method is calculated, using RF and 5NN as base classifiers. The learning time of the classifiers themselves are also included for comparison. The number of annotators is varied between 3 and 6 to see its effect in the execution time. Also, apart from considering all 6 class labels from the dataset, the execution time considering only a half of the class labels (3) is also computed to understand the impact of the size of the class variable.

The results of this scalability test can be observed in Figure 6. The numbers of instances, classes and annotators affect the running times of both of our methods, so does the choice of the classifier. The running time of SL-C is always lower than that of JL, which seems to grow exponentially when employing RF and linearly in the case of 5NN. When doubling the number of annotators, the running time of our methods has only a small increase. However, when the number of classes is reduced to a half, the difference is notorious, reducing the running time nearly 8 or 10 times for JL-C with RF, and about 4 times for SL-C with 5NN.

References

[1] L. Liu and T. Dietterich, "A conditional multinomial mixture model for superset label learning," *Advances in neural information processing systems*, vol. 25, 2012.



(a) Dataset *arrhythmia*

(b) Dataset *dermatology*

(c) Dataset *glass*

(d) Dataset *pendigits*

(e) Dataset *satimage*

(f) Dataset *segment*

(g) Dataset *svmguide4*

(h) Dataset *vehicle*

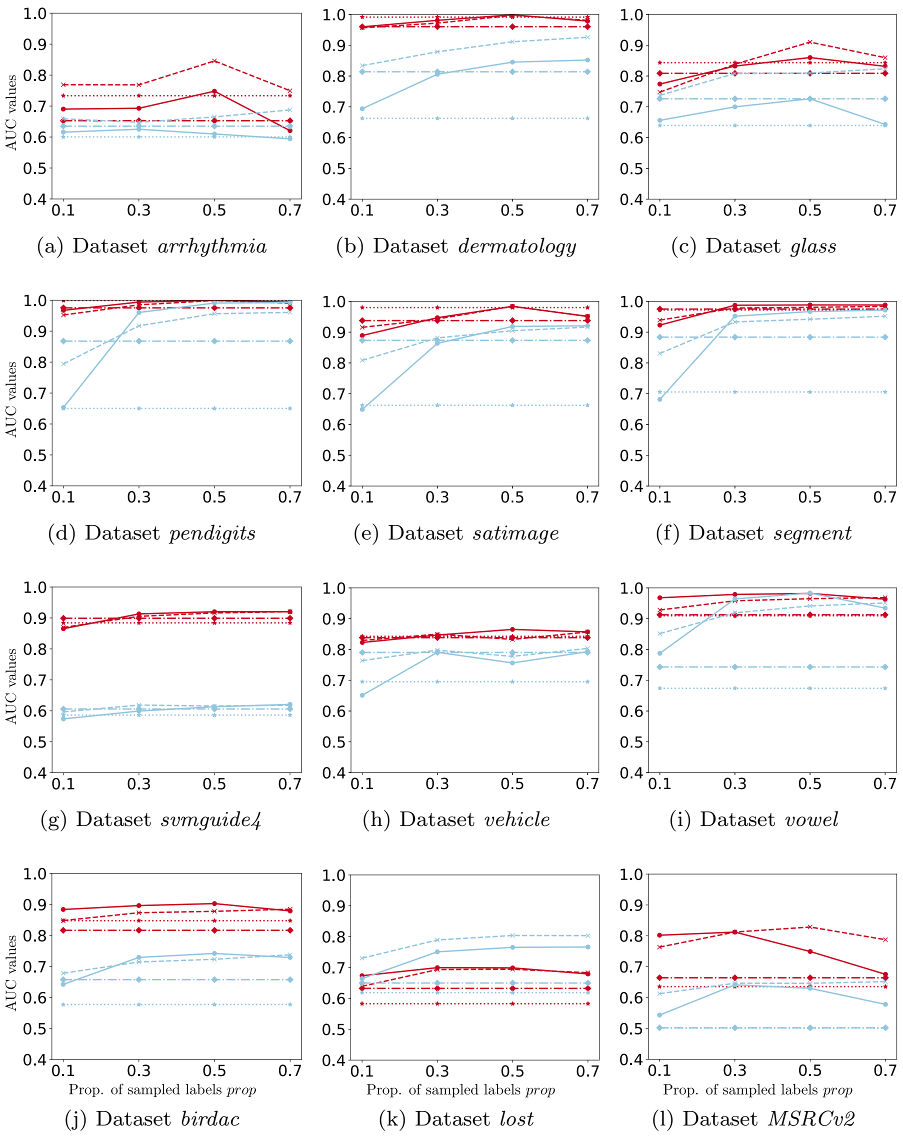
(i) Dataset *vowel*

(j) Dataset *birdac*

(k) Dataset *lost*

(l) Dataset *MSRCv2*

Figure 4: Experimental results throughout different values of the parameter m (number of annotators), in terms of AUC metric, within different datasets (subplots). Results with classifiers RF and 5NN are displayed in dark blue and light blue colour, respectively. A different line style and marker is used for each method (SL-C, JL-C, RAY, DS). The rest of generative parameters are fixed to $\beta = 5$ and $prop = 0.5$.



(a) Dataset *arrhythmia*

(b) Dataset *dermatology*

(c) Dataset *glass*

(d) Dataset *pendigits*

(e) Dataset *satimage*

(f) Dataset *segment*

(g) Dataset *svmguide4*

(h) Dataset *vehicle*

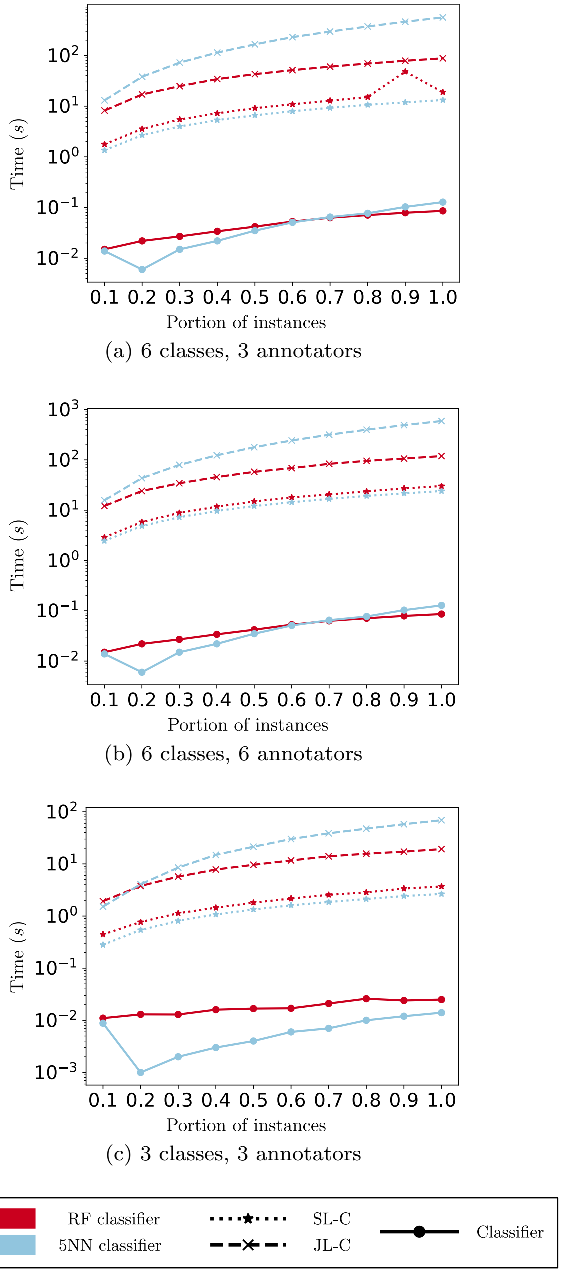
(i) Dataset *vowel*

(j) Dataset *birdac*

(k) Dataset *lost*

(l) Dataset *MSRCv2*

Figure 5: Experimental results throughout different values of the parameter *prop* (flexibility of the annotators), in terms of AUC metric, within different datasets (subplots). Results with classifiers RF and 5NN are displayed in dark blue and light blue colour, respectively. A different line style and marker is used for each method (SL-C, JL-C, RAY, DS). The rest of generative parameters are fixed to $\beta = 5$ and $m = 5$.



(a) 6 classes, 3 annotators

(b) 6 classes, 6 annotators

(c) 3 classes, 3 annotators

Figure 6: Scalability test with our two methods, SL-C and JL-C, and the classifiers RF and 5NN, throughout different portions of the complete dataset. Running time is measured in seconds and shown in logarithmic scale in the Y axis. Results with classifiers RF and 5NN are displayed in red and light blue colour, respectively. Different line styles and markers are used for each method. Each subfigure shows the performance with varying numbers of classes and annotators.