Progress in Artificial Intelligence manuscript No.
(will be inserted by the editor)

Supplementary material for: "*A framework for evaluation in learning from label proportions*"

Jerónimo Hernández-González

Last update: 04/04/2019

Jerónimo Hernández-González
IIIA-CSIC, Campus UAB, 08913 Bellaterra, Catalonia, Spain. Tel.: +34 935 809 570. E-mail: momo@iiia.csic.es

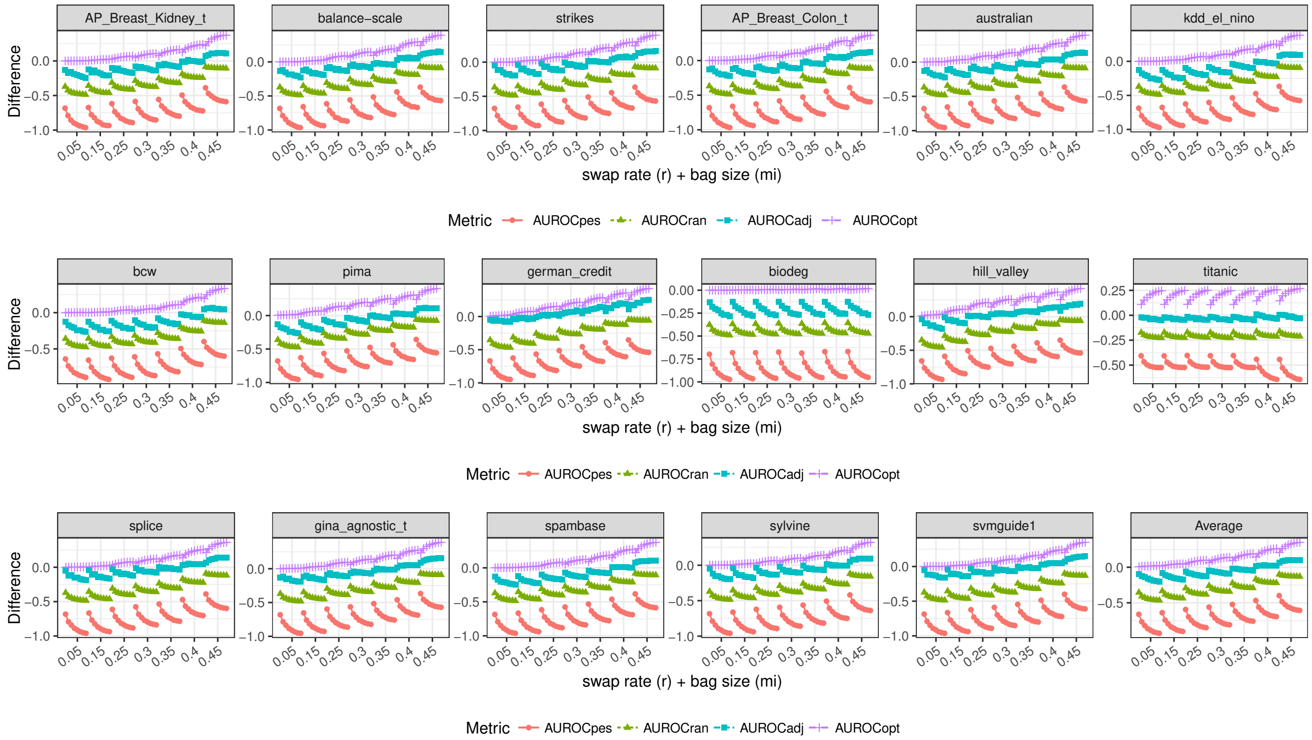

Fig. 1 Experimental results with **corrupted training sets** in different datasets and averaged over all of them (bottom-right figure). Using **Random Forest classifiers**, each figure shows the difference between the **AUC-ROC** estimation using the original completely labeled data and the AUC-ROC values obtained from the proposed different approximations (pessimistic, random, adjusted and optimistic)to the TP count. To simulate different experimental conditions, an increasingly worse classifier is induced by corrupting the r percentage of the training data labels (with $r \in \{5\%, 15\%, 25\%, 30\%, 35\%, 40\%, 45\%\}$; shown in the figures by means of different lines) and the LLP settings in the validation partition is simulated with bags of increasing size, m_i (with $m_i = \{4, 7, 10, 15, 20, 30, 40, 50\}$; represented in the figures by the different points of each line).

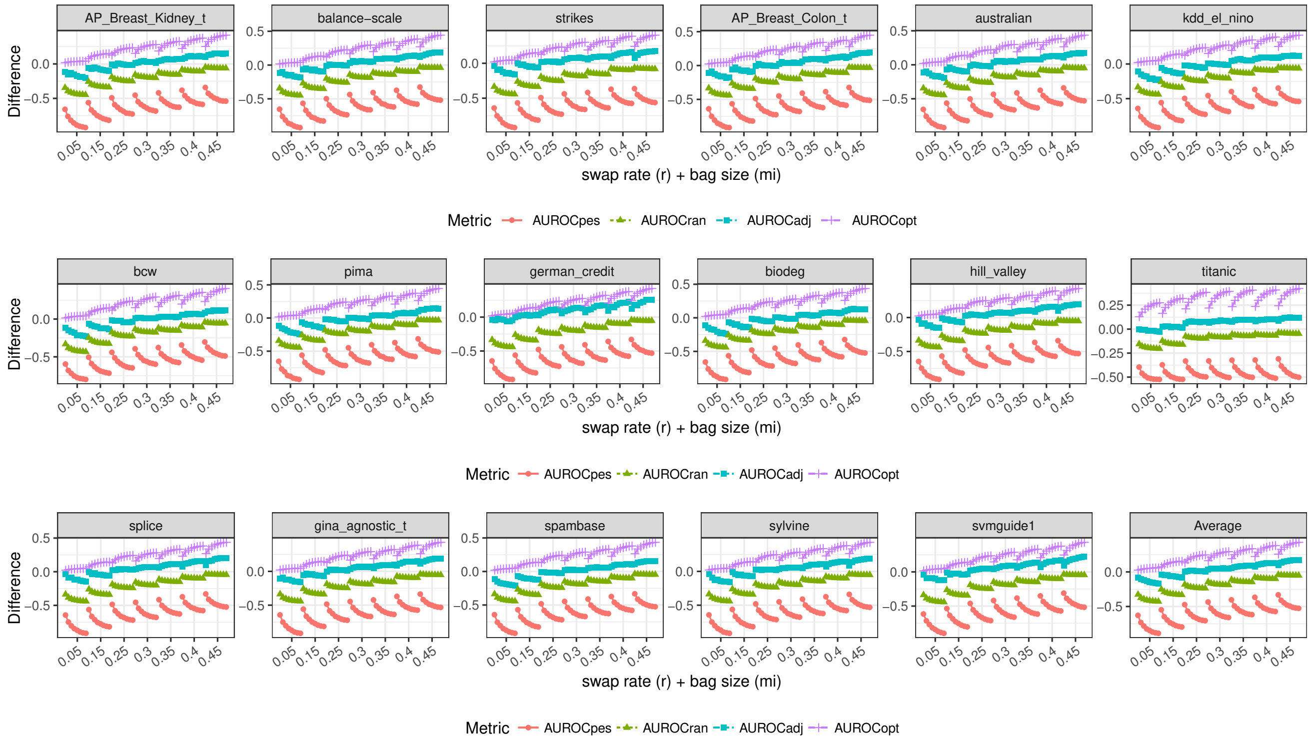

Fig. 2 Experimental results with **corrupted predictions** in different datasets and averaged over all of them (bottom-right figure). Using **Random Forest classifiers**, each figure shows the difference between the **AUC-ROC** estimation using the original completely labeled data and the AUC-ROC values obtained from the proposed different approximations (pessimistic, random, adjusted and optimistic)to the TP count. To simulate different experimental conditions, an increasingly worse classifier is induced by corrupting the r percentage of the training data labels (with $r \in \{5\%, 15\%, 25\%, 30\%, 35\%, 40\%, 45\%\}$; shown in the figures by means of different lines) and the LLP settings in the validation partition is simulated with bags of increasing size, m_i (with $m_i = \{4, 7, 10, 15, 20, 30, 40, 50\}$; represented in the figures by the different points of each line).

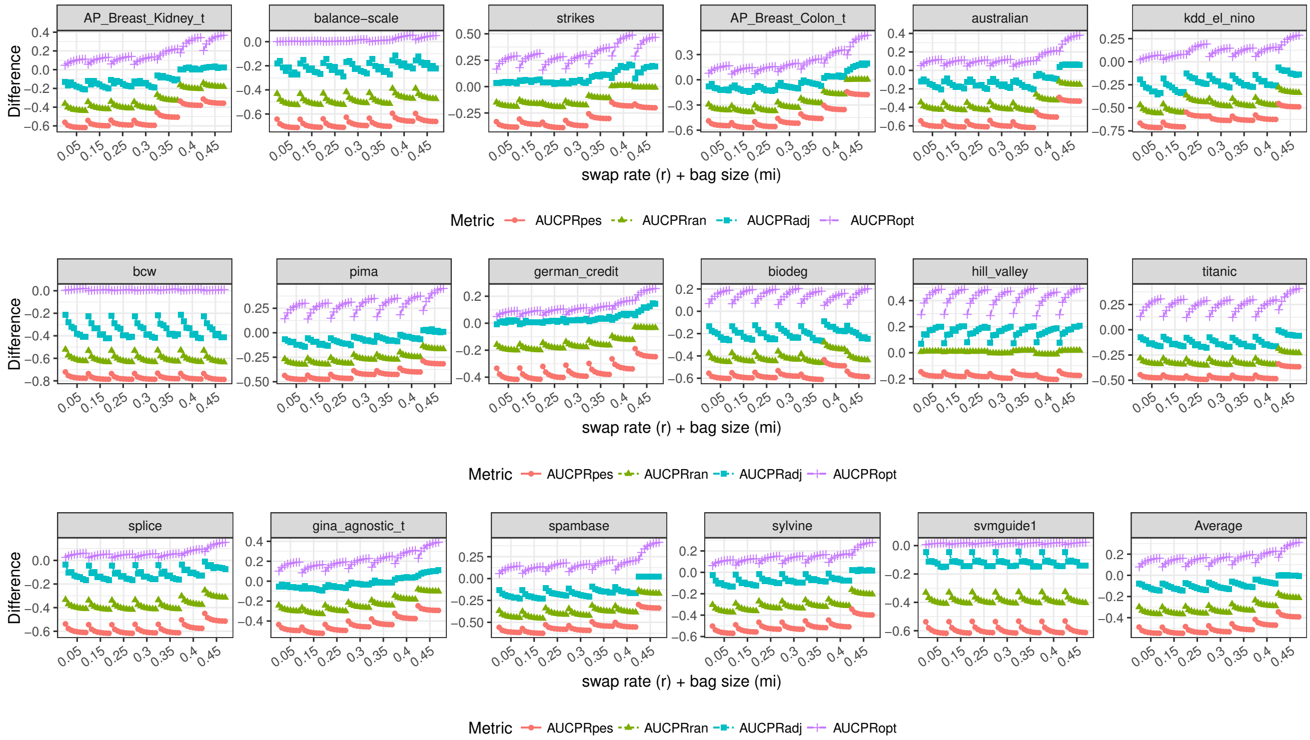

Fig. 3 Experimental results with **corrupted training sets** in different datasets and averaged over all of them (bottom-right figure). Using **Naive Bayes classifiers**, each figure shows the difference between the **AUC-PR** estimation using the original completely labeled data and the AUC-PR values obtained from the proposed different approximations (pessimistic, random, adjusted and optimistic)to the TP count. To simulate different experimental conditions, an increasingly worse classifier is induced by corrupting the r percentage of the training data labels (with $r \in \{5\%, 15\%, 25\%, 30\%, 35\%, 40\%, 45\%\}$; shown in the figures by means of different lines) and the LLP settings in the validation partition is simulated with bags of increasing size, m_i (with $m_i = \{4, 7, 10, 15, 20, 30, 40, 50\}$; represented in the figures by the different points of each line).

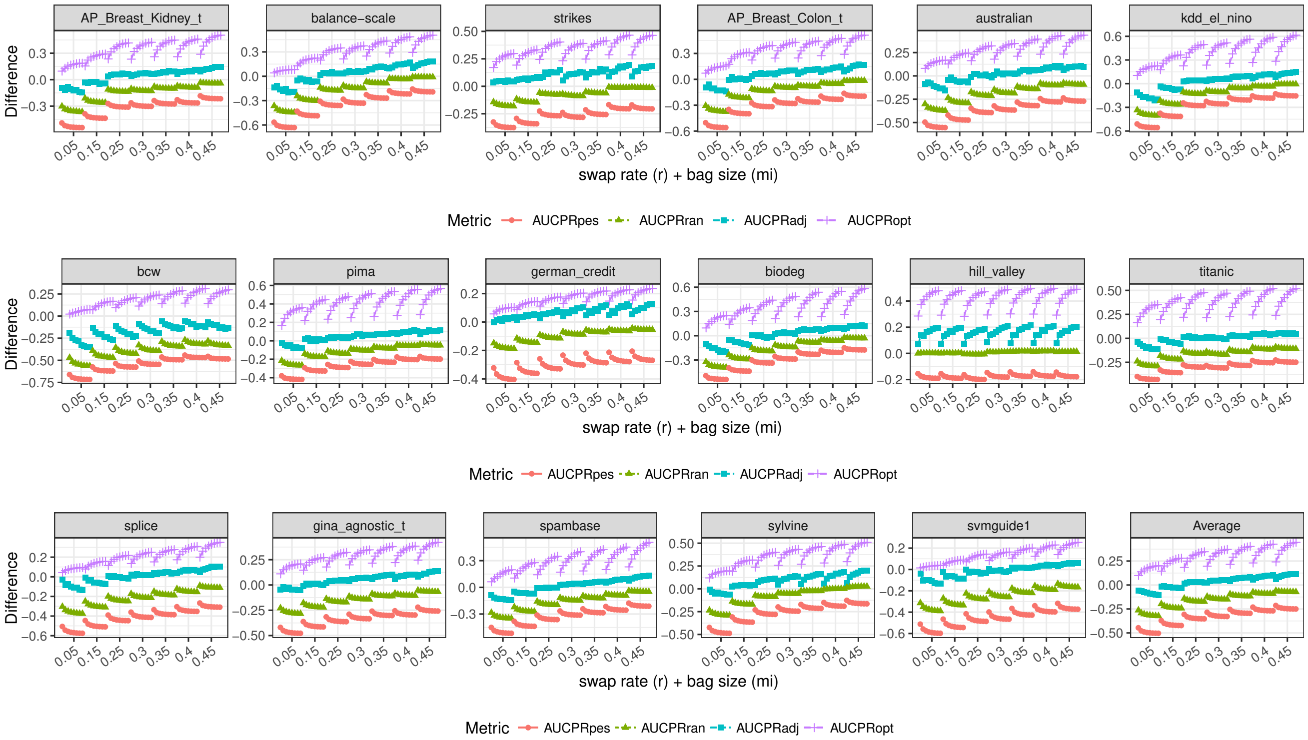

Fig. 4 Experimental results with **corrupted predictions** in different datasets and averaged over all of them (bottom-right figure). Using **Naive Bayes classifiers**, each figure shows the difference between the **AUC-PR** estimation using the original completely labeled data and the AUC-PR values obtained from the proposed different approximations (pessimistic, random, adjusted and optimistic)to the TP count. To simulate different experimental conditions, an increasingly worse classifier is induced by corrupting the r percentage of the training data labels (with $r \in \{5\%, 15\%, 25\%, 30\%, 35\%, 40\%, 45\%\}$; shown in the figures by means of different lines) and the LLP settings in the validation partition is simulated with bags of increasing size, m_i (with $m_i = \{4, 7, 10, 15, 20, 30, 40, 50\}$; represented in the figures by the different points of each line).

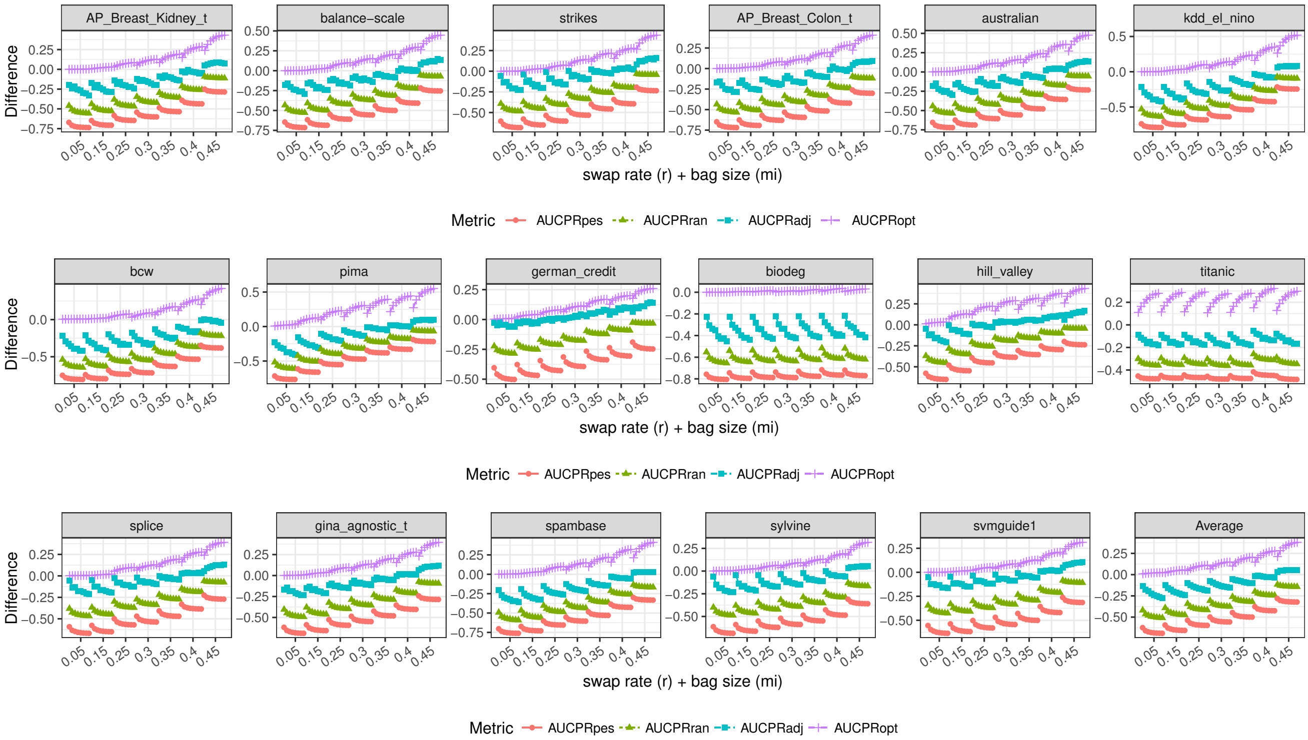

Fig. 5 Experimental results with **corrupted training sets** in different datasets and averaged over all of them (bottom-right figure). Using **Random Forest classifiers**, each figure shows the difference between the **AUC-PR** estimation using the original completely labeled data and the AUC-PR values obtained from the proposed different approximations (pessimistic, random, adjusted and optimistic)to the TP count. To simulate different experimental conditions, an increasingly worse classifier is induced by corrupting the r percentage of the training data labels (with $r \in \{5\%, 15\%, 25\%, 30\%, 35\%, 40\%, 45\%\}$; shown in the figures by means of different lines) and the LLP settings in the validation partition is simulated with bags of increasing size, m_i (with $m_i = \{4, 7, 10, 15, 20, 30, 40, 50\}$; represented in the figures by the different points of each line).

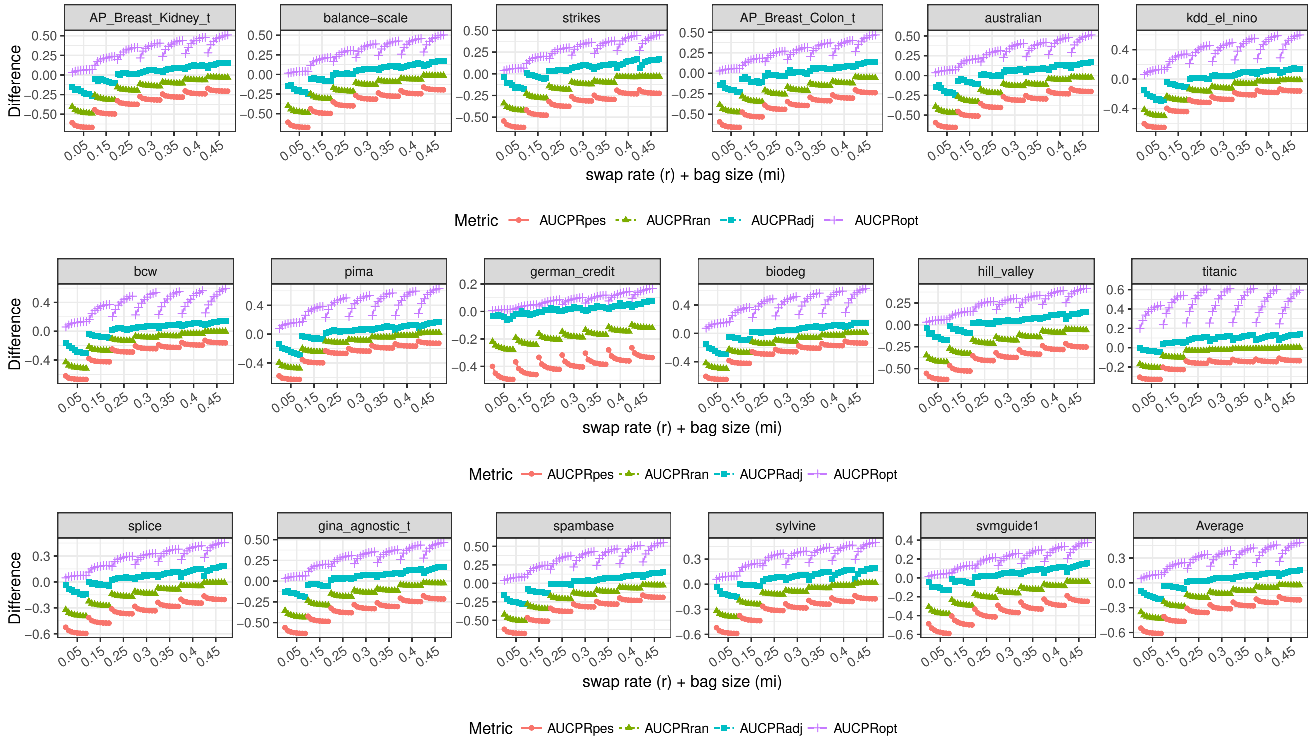

Fig. 6 Experimental results with **corrupted predictions** in different datasets and averaged over all of them (bottom-right figure). Using **Random Forest classifiers**, each figure shows the difference between the **AUC-PR** estimation using the original completely labeled data and the AUC-PR values obtained from the proposed different approximations (pessimistic, random, adjusted and optimistic)to the TP count. To simulate different experimental conditions, an increasingly worse classifier is induced by corrupting the r percentage of the training data labels (with $r \in \{5\%, 15\%, 25\%, 30\%, 35\%, 40\%, 45\%\}$; shown in the figures by means of different lines) and the LLP settings in the validation partition is simulated with bags of increasing size, m_i (with $m_i = \{4, 7, 10, 15, 20, 30, 40, 50\}$; represented in the figures by the different points of each line).

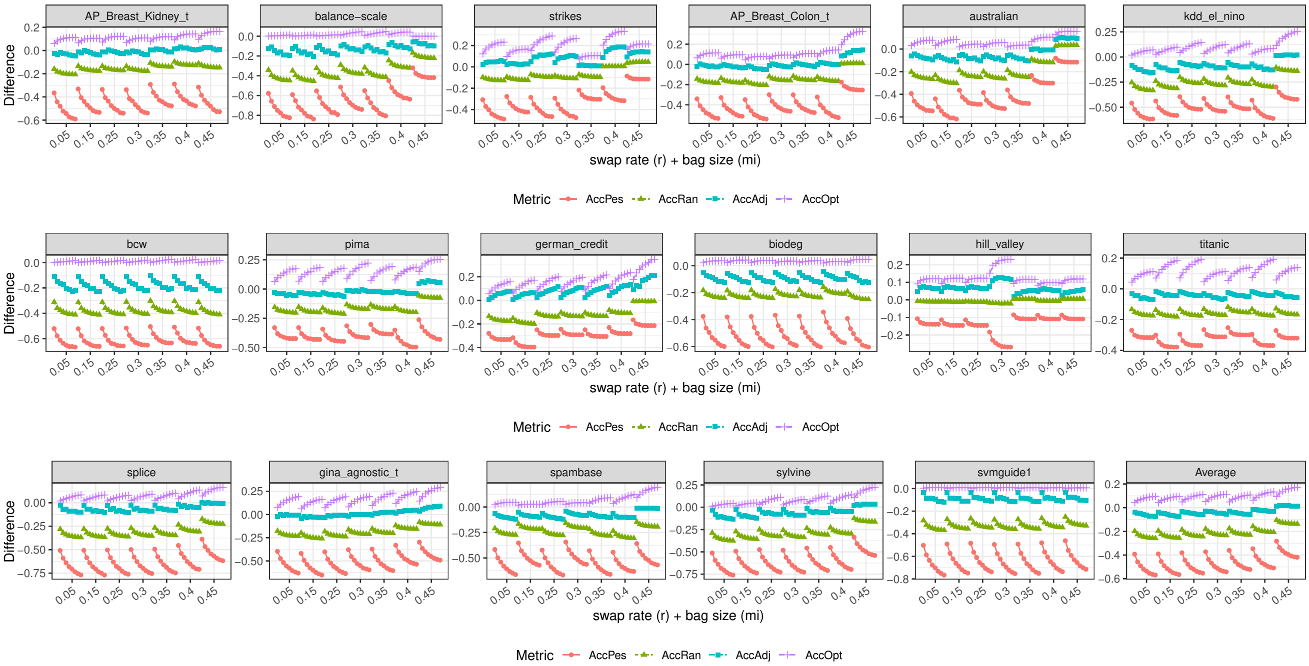

Fig. 7 Experimental results with **corrupted training sets** in different datasets and averaged over all of them (bottom-right figure). Using **Naive Bayes classifiers**, each figure shows the difference between the **Accuracy** estimation using the original completely labeled data and the Accuracy values obtained from the proposed different approximations (pessimistic, random, adjusted and optimistic)to the TP count. To simulate different experimental conditions, an increasingly worse classifier is induced by corrupting the r percentage of the training data labels (with $r \in \{5\%, 15\%, 25\%, 30\%, 35\%, 40\%, 45\%\}$; shown in the figures by means of different lines) and the LLP settings in the validation partition is simulated with bags of increasing size, m_i (with $m_i = \{4, 7, 10, 15, 20, 30, 40, 50\}$; represented in the figures by the different points of each line).

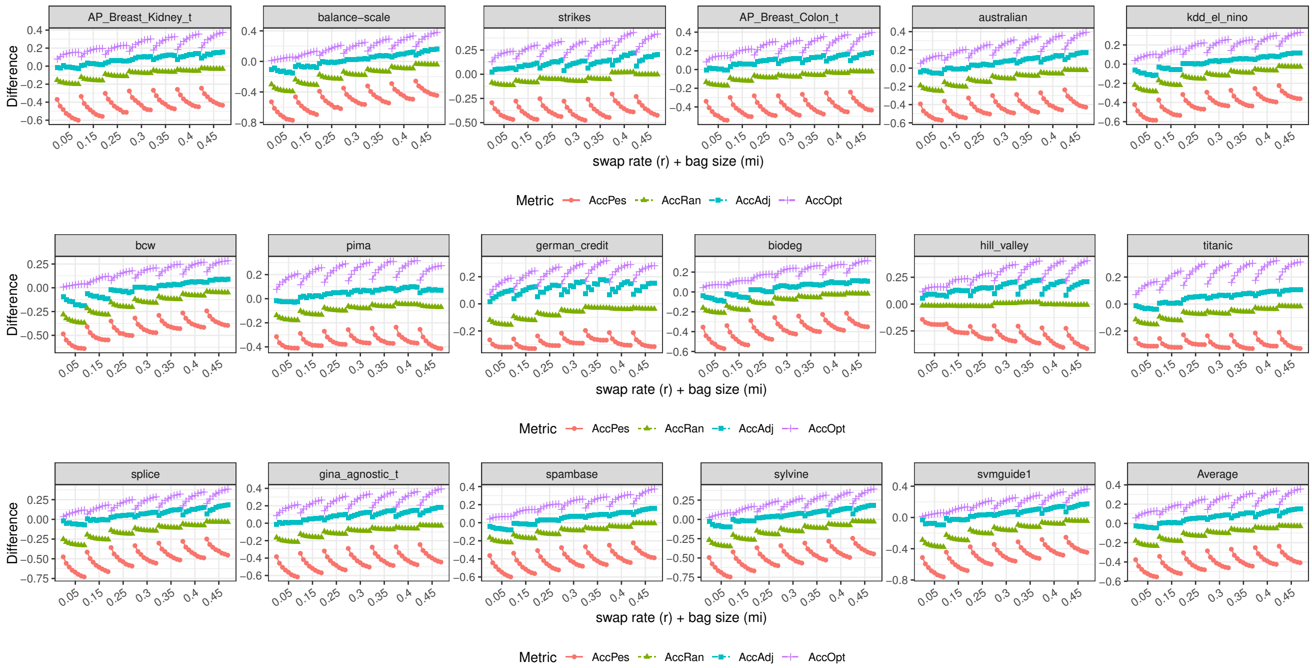

Fig. 8 Experimental results with **corrupted predictions** in different datasets and averaged over all of them (bottom-right figure). Using **Naive Bayes classifiers**, each figure shows the difference between the **Accuracy** estimation using the original completely labeled data and the Accuracy values obtained from the proposed different approximations (pessimistic, random, adjusted and optimistic)to the TP count. To simulate different experimental conditions, an increasingly worse classifier is induced by corrupting the r percentage of the training data labels (with $r \in \{5\%, 15\%, 25\%, 30\%, 35\%, 40\%, 45\%\}$; shown in the figures by means of different lines) and the LLP settings in the validation partition is simulated with bags of increasing size, m_i (with $m_i = \{4, 7, 10, 15, 20, 30, 40, 50\}$; represented in the figures by the different points of each line).

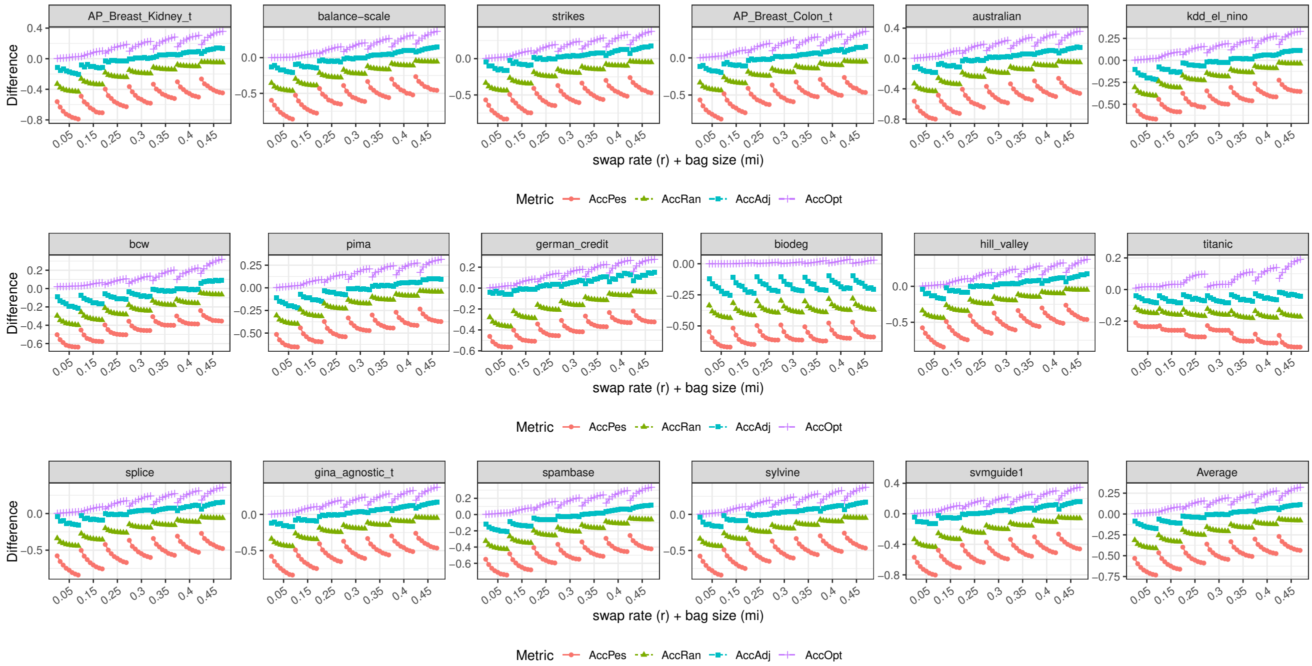

Fig. 9 Experimental results with **corrupted training sets** in different datasets and averaged over all of them (bottom-right figure). Using **Random Forest classifiers**, each figure shows the difference between the **Accuracy** estimation using the original completely labeled data and the Accuracy values obtained from the proposed different approximations (pessimistic, random, adjusted and optimistic)to the TP count. To simulate different experimental conditions, an increasingly worse classifier is induced by corrupting the r percentage of the training data labels (with $r \in \{5\%, 15\%, 25\%, 30\%, 35\%, 40\%, 45\%\}$; shown in the figures by means of different lines) and the LLP settings in the validation partition is simulated with bags of increasing size, m_i (with $m_i = \{4, 7, 10, 15, 20, 30, 40, 50\}$; represented in the figures by the different points of each line).

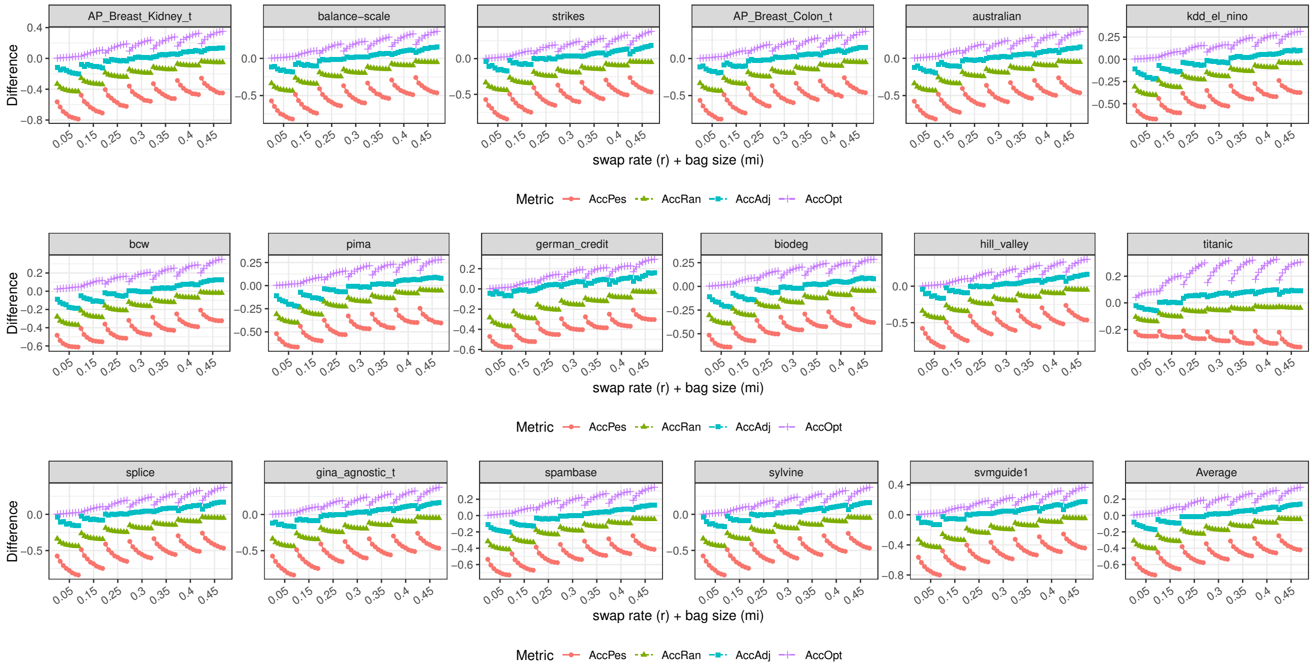

Fig. 10 Experimental results with **corrupted predictions** in different datasets and averaged over all of them (bottom-right figure). Using **Random Forest classifiers**, each figure shows the difference between the **Accuracy** estimation using the original completely labeled data and the Accuracy values obtained from the proposed different approximations (pessimistic, random, adjusted and optimistic)to the TP count. To simulate different experimental conditions, an increasingly worse classifier is induced by corrupting the r percentage of the training data labels (with $r \in \{5\%, 15\%, 25\%, 30\%, 35\%, 40\%, 45\%\}$; shown in the figures by means of different lines) and the LLP settings in the validation partition is simulated with bags of increasing size, m_i (with $m_i = \{4, 7, 10, 15, 20, 30, 40, 50\}$; represented in the figures by the different points of each line).