

Hitting the Target: Stopping Active Learning at the Cost-Based Optimum Supplementary Information

Contents

1	Metric Examples	1
2	Additional SVM Results	2
3	Random Forest Results	5
4	Neural Network Results	12
5	SVM Class Imbalance	18

1 Metric Examples

To illustrate how the metrics used in the SC discussed above behave we plot values from an exemplary run in Fig. 1. Each plot shows the *metric* as well as the true accuracy of the classifier. In a real AL scenario this would not be available, but we keep a separate hold-out set on which to evaluate accuracy so we can compare SC performance. Thresholds (if the condition uses one) are shown with dashed horizontal lines. We observe that some metrics seem to be correlated with accuracy (STABILIZING PREDICTIONS) while others, such as SSNCUT, are not obviously connected.

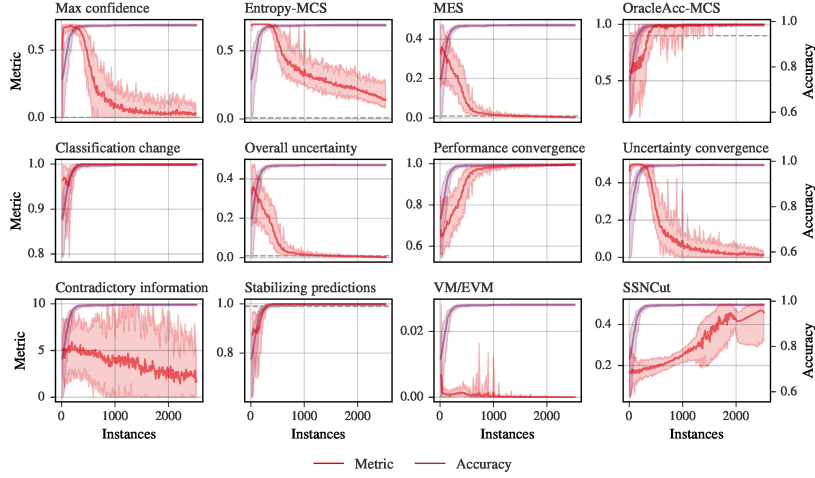


Figure 1: Exemplary metric values during an AL process. The ground-truth (but in AL unknown) accuracy is provided for comparison. These results are from the spamassassin dataset using a linear SVM. The 2.5% to 97.5% percentiles are shown in filled colour. Thresholds, if used by the condition, are displayed as a grey dashed line

2 Additional SVM Results

In Figures 2 and 3 we show our results when including SC with out penalty, and excluding them completely when they failed to stop respectively.

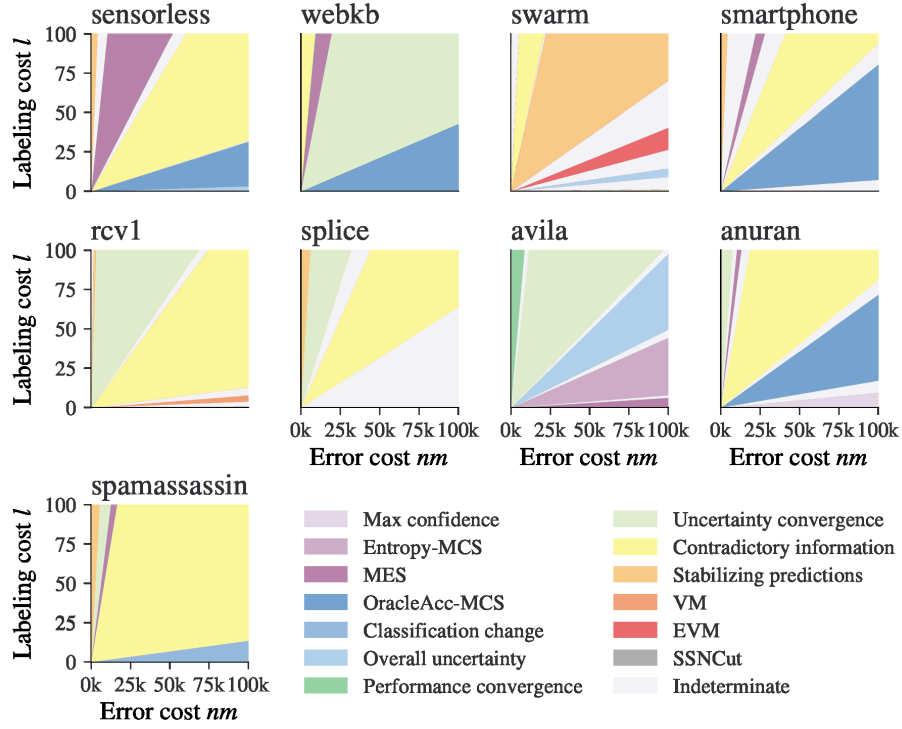


Figure 2: The regions in which each criterion is cost-optimal by dataset on SVMs. Criteria that failed to stop were included without penalty. Light grey shows regions in which no SC was statistically dominant

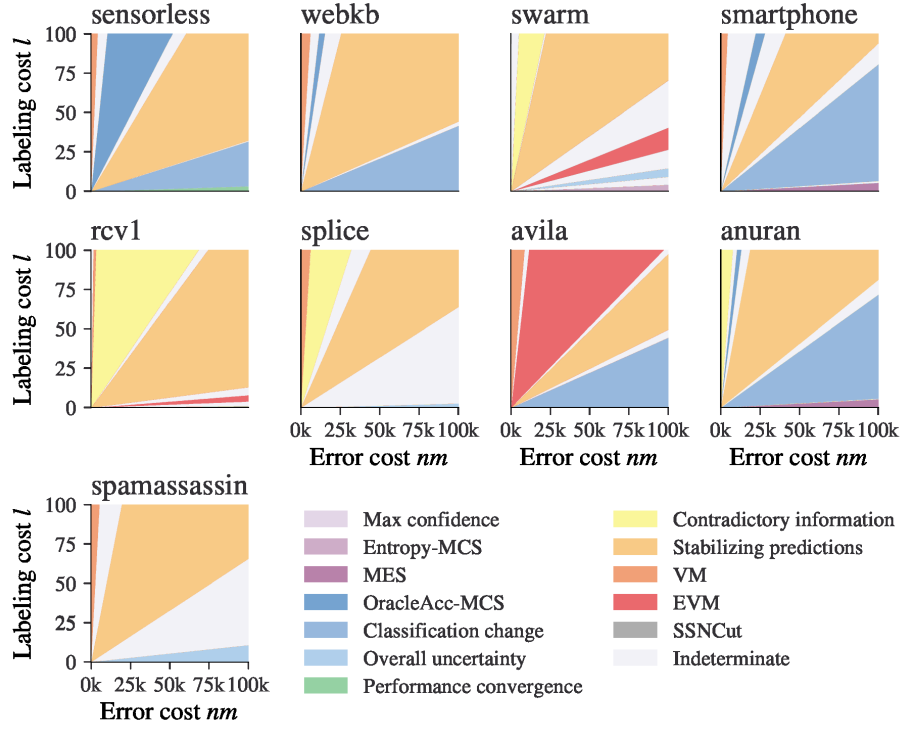


Figure 3: The regions in which each criterion is cost-optimal by dataset on SVMs. Criteria that failed to stop were excluded. Light grey shows regions in which no SC was statistically dominant

3 Random Forest Results

We present our results from evaluating stopping conditions using random forests. All figures are generated in the same manner as their SVM counterparts above. SSNCUT was not evaluated under random forests as it was motivated only for SVMs. Figure 4 shows the stopping points of the conditions. Figures 5, 6, 7, and 8 show the cost-optimal SC over a range of values of the cost parameters n , m , and l . Figure 9 and 10 show the rankings of SC in our two example scenarios. Figure 11 shows the number of times each criterion stops on each dataset and the average correlations between their metrics and the classifier’s accuracy. Table 2 shows the correlations between the SC’s metrics and the classifier’s accuracy.

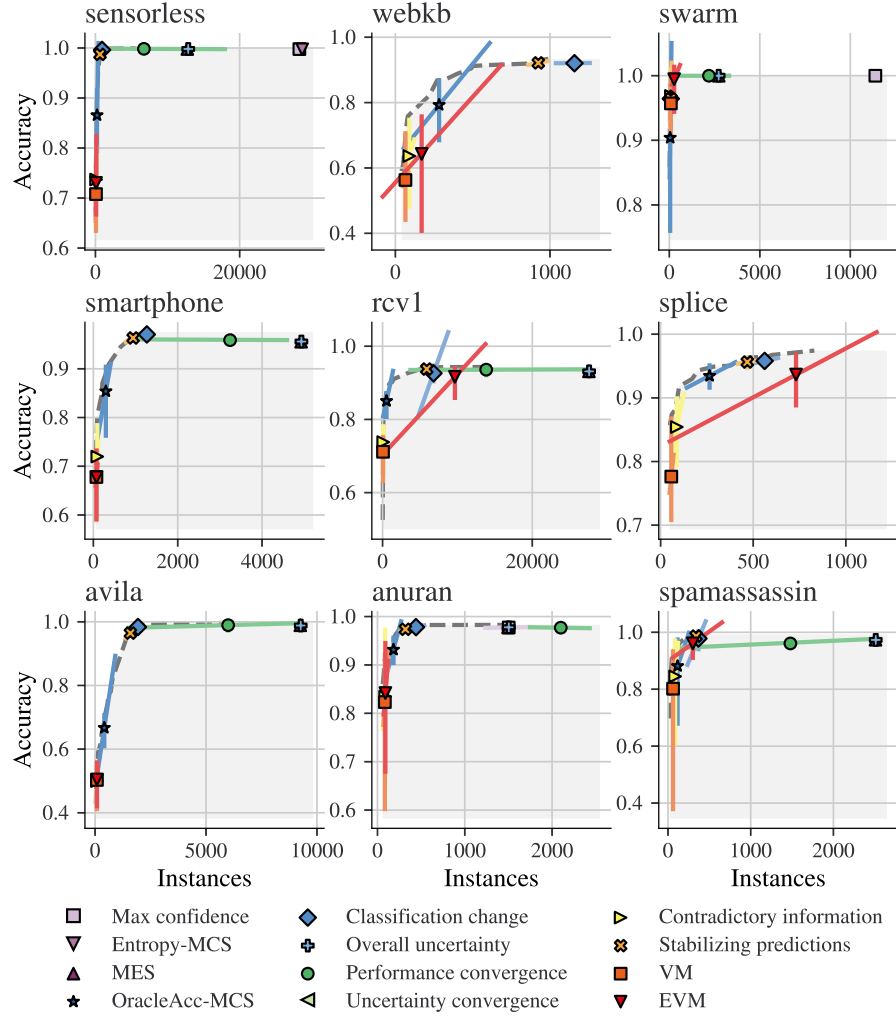


Figure 4: The accuracy and number of instances used by SC when evaluated on a random forest classifier. The dashed grey line shows the Pareto frontier. Error bars were computed from 2.5% and 97.5% percentiles and transformed by PCA

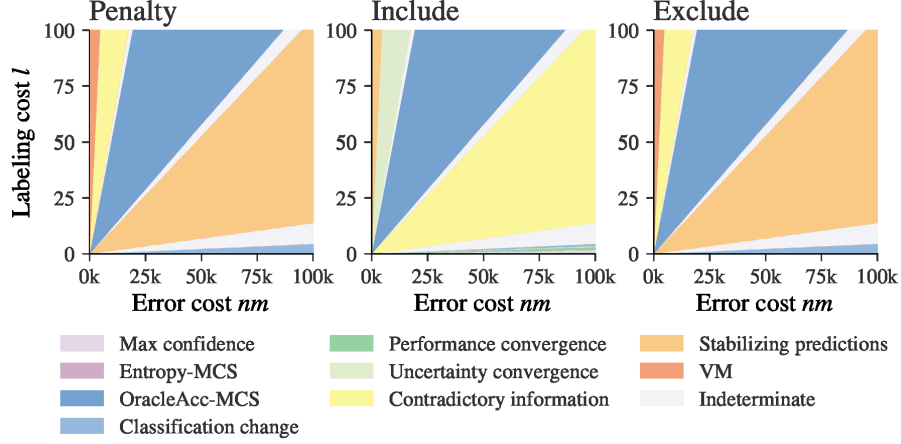


Figure 5: The regions in which each criterion was cost-optimal on random forests. Each plot shows different treatment of SC which failed to stop. In penalty they were penalized with the worst instances and accuracy, while in include/exclude they were included without penalty or excluded. Light grey shows regions in which no SC was statistically dominant

Table 1: Pearson correlation coefficients on random forests

Criterion	Correlation
Max confidence	0.39 ± 0.06
Entropy-MCS	0.36 ± 0.06
MES	0.56 ± 0.05
OracleAcc-MCS	0.63 ± 0.04
Classification change	0.83 ± 0.03
Overall uncertainty	0.56 ± 0.05
Performance convergence	0.55 ± 0.05
Uncertainty convergence	0.40 ± 0.05
Contradictory information	0.05 ± 0.02
Stabilizing predictions	0.90 ± 0.03
VM	0.37 ± 0.10
EVM	0.37 ± 0.10

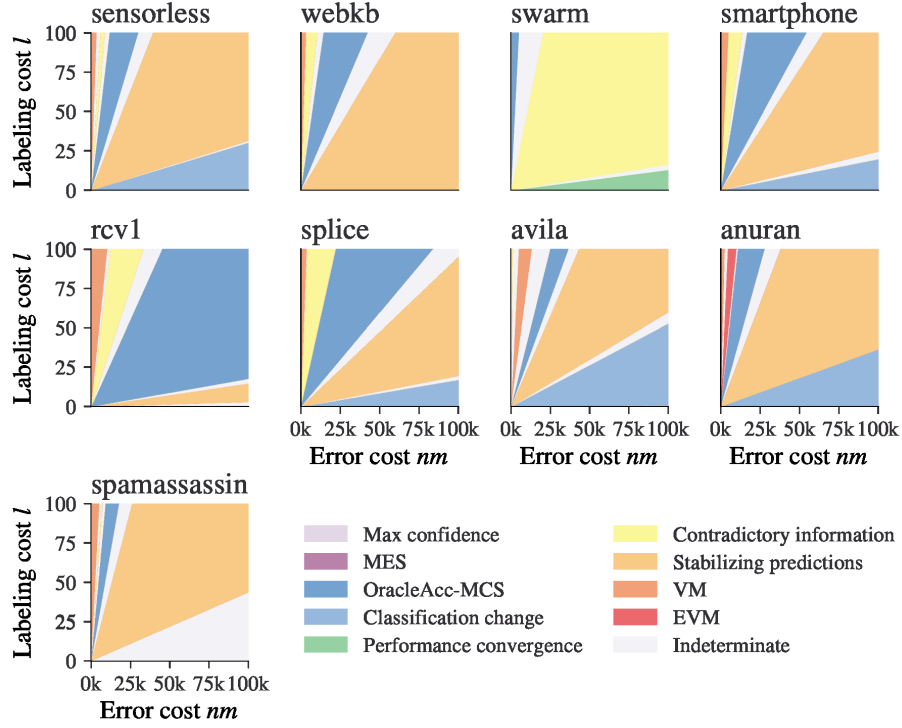


Figure 6: The regions in which each criterion is cost-optimal by dataset on random forests. Criteria were penalised for each split of each dataset on which they failed to stop. Light grey shows regions in which no SC was statistically dominant

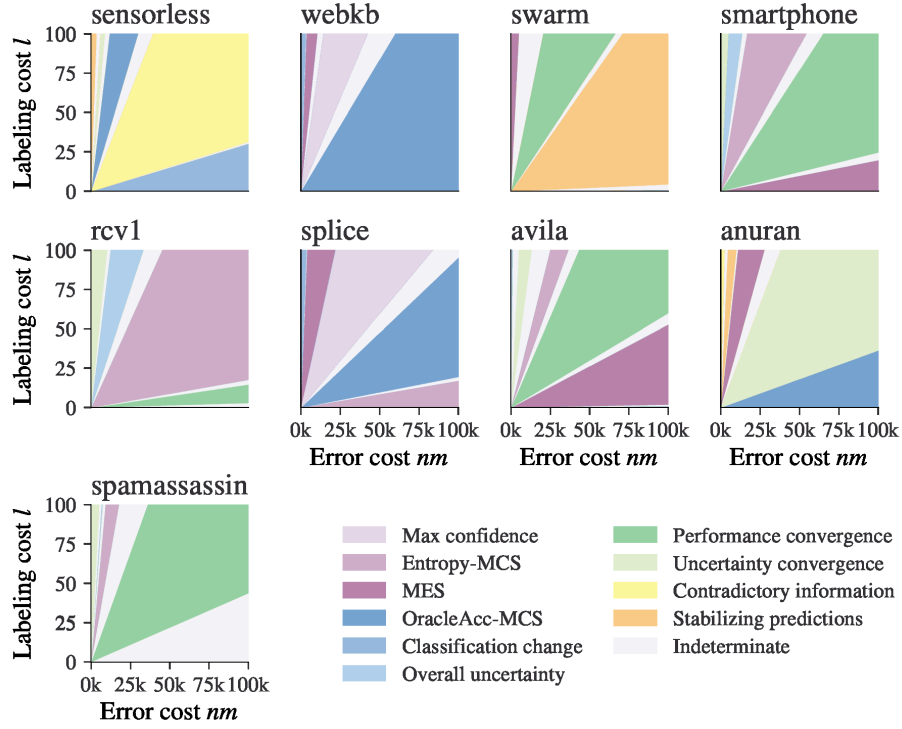


Figure 7: The regions in which each criterion is cost-optimal by dataset on random forests. Criteria that failed to stop were included without penalty. Light grey shows regions in which no SC was statistically dominant

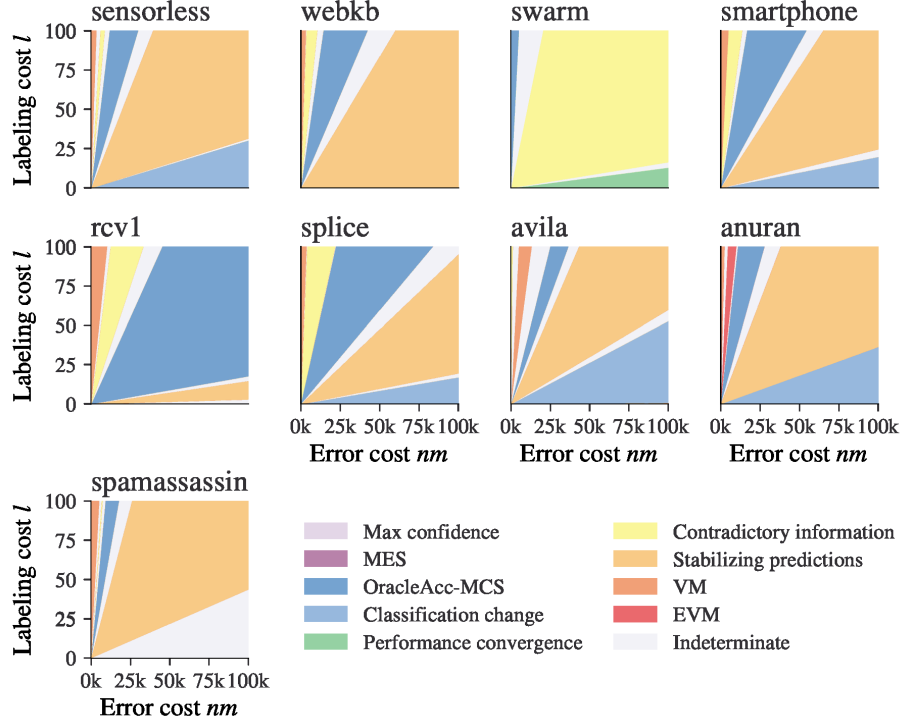


Figure 8: The regions in which each criterion is cost-optimal by dataset on random forests. Criteria that failed to stop were excluded. Light grey shows regions in which no SC was statistically dominant

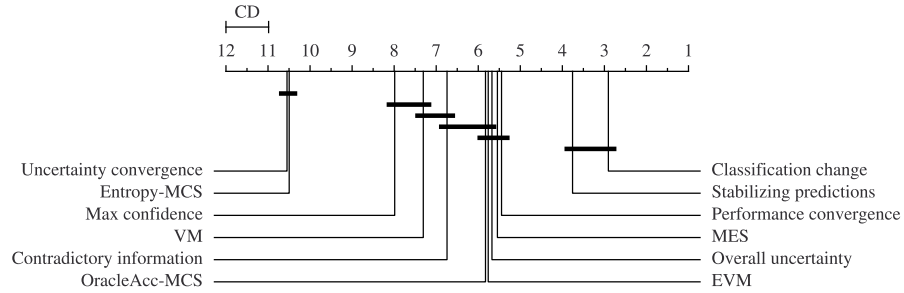


Figure 9: The relative cost of SC for an example scenario classifying mammo-grams using a random forest model

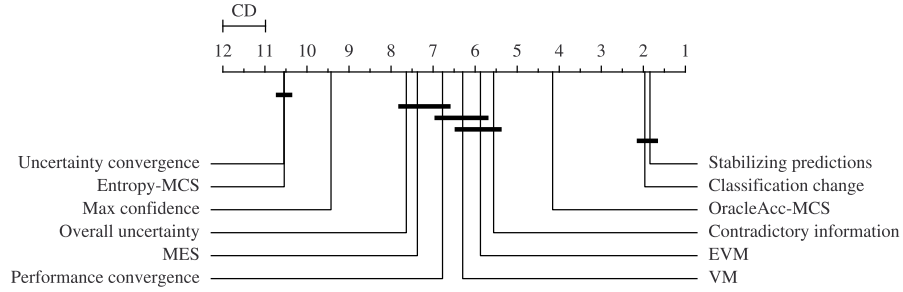


Figure 10: The relative cost of SC for an example scenario selecting potential customers using a random forest model

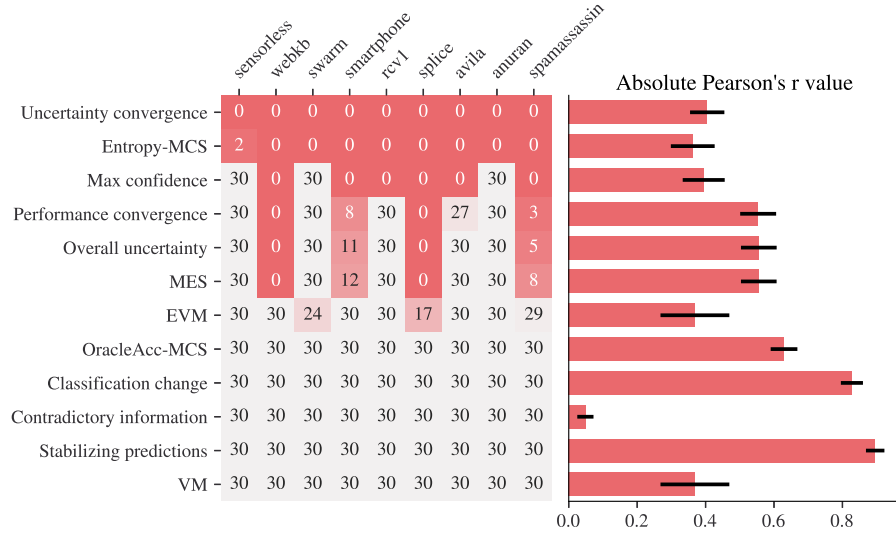


Figure 11: Left: the number of times each criterion stopped on each dataset using random forests out of 30 splits. Right: the Pearson correlation coefficient between each SC's metric and the classifier's accuracy. The black shows the standard error of the mean across datasets

4 Neural Network Results

We present our results from evaluating stopping conditions using a single hidden layer NN. All figures are generated in the same manner as their SVM counterparts above. SSNCUT was not evaluated under NNs as it was motivated only for SVMs. rcv1 was not used in our NN evaluations as the large size and high dimensionality made it computationally expensive. Figure 12 shows the stopping points of the conditions. Figures 13, 14, 15, and 16 show the cost-optimal SC over a range of values of the cost parameters n , m , and l . Figure 17, and 18 show the rankings of SC in our two example scenarios. Figure 19 shows the number of times each criterion stops on each dataset and the average correlations between their metrics and the classifier’s accuracy. Table 2 shows the correlations between the SC’s metrics and the classifier’s accuracy.

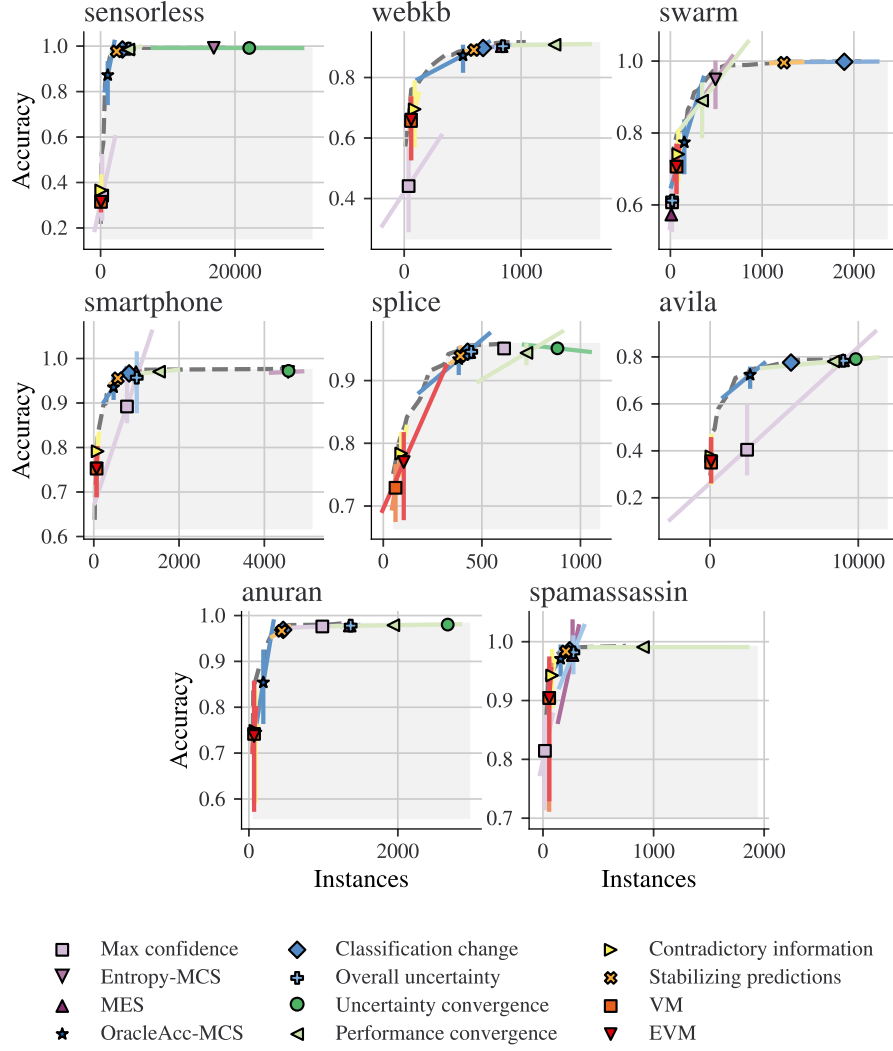


Figure 12: The accuracy and number of instances used by SC when evaluated on an NN. The dashed grey line shows the Pareto frontier. Error bars were computed from 2.5% and 97.5% percentiles and transformed by PCA

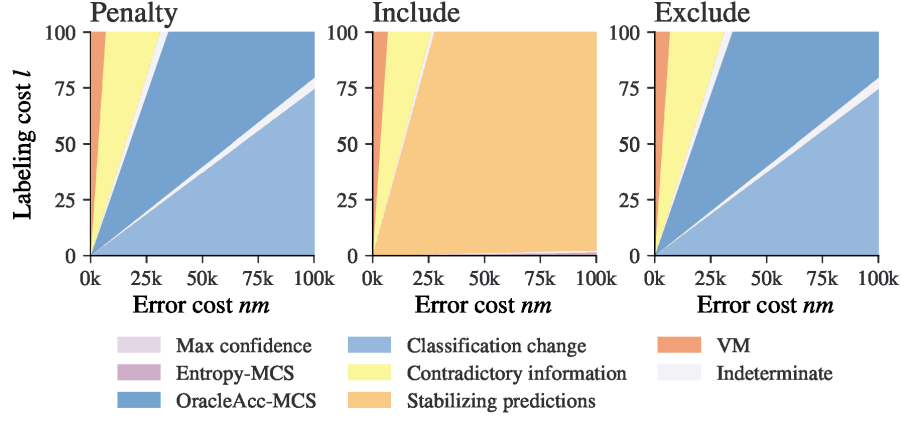


Figure 13: The regions in which each criterion was cost-optimal on an NN. Each plot shows different treatment of SC which failed to stop. In penalty they were penalized with the worst instances and accuracy, while in include/exclude they were included without penalty or excluded. Light grey shows regions in which no SC was statistically dominant

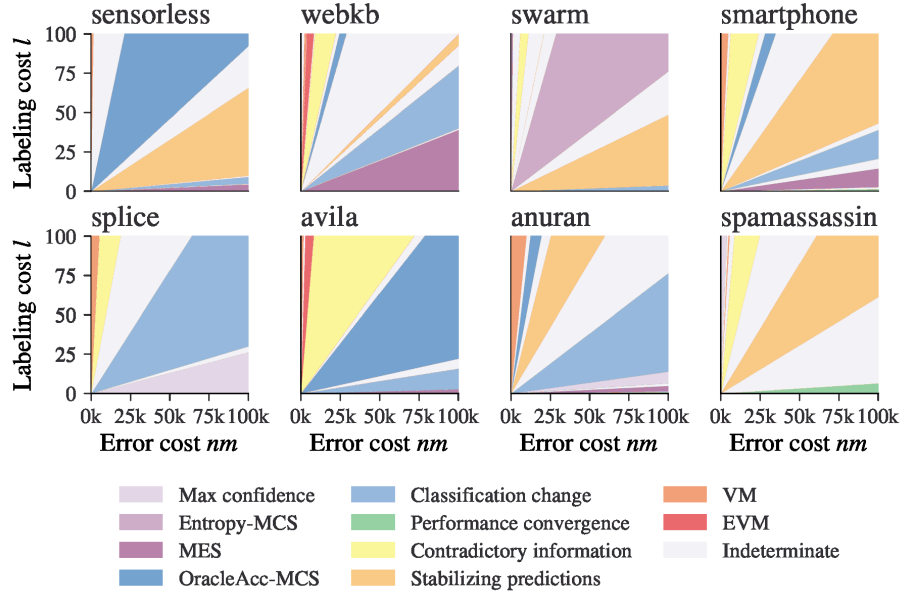


Figure 14: The regions in which each criterion is cost-optimal by dataset on an NN. Criteria were penalised for each split of each dataset on which they failed to stop

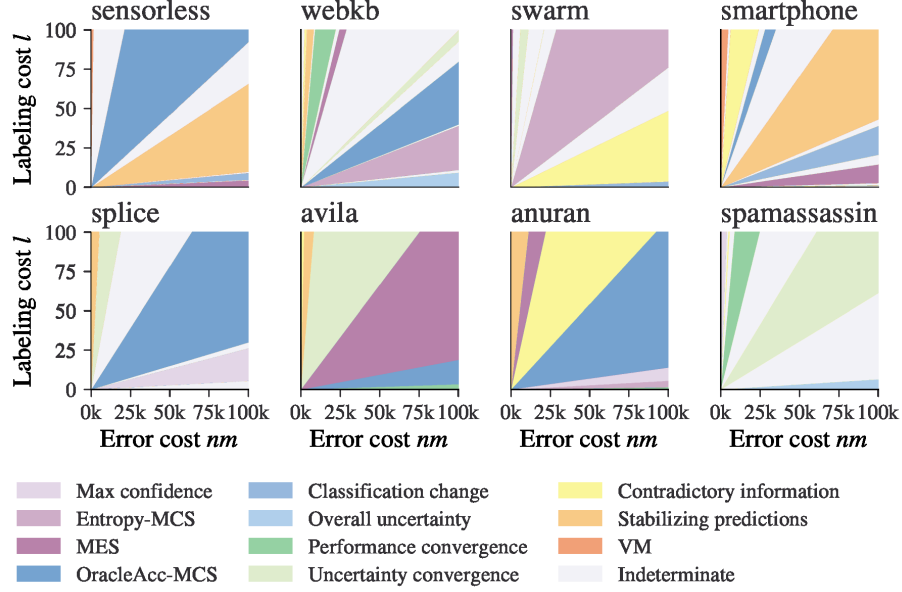


Figure 15: The regions in which each criterion is cost-optimal by dataset on an NN. Criteria that failed to stop were included without penalty. Light grey shows regions in which no SC was statistically dominant

Table 2: Pearson correlation coefficients on NNs

Criterion	Correlation
Max confidence	0.51 ± 0.04
Entropy-MCS	0.61 ± 0.03
MES	0.85 ± 0.03
OracleAcc-MCS	0.65 ± 0.02
Classification change	0.89 ± 0.02
Overall uncertainty	0.81 ± 0.03
Performance convergence	0.84 ± 0.03
Uncertainty convergence	0.58 ± 0.02
Contradictory information	0.05 ± 0.03
Stabilizing predictions	0.92 ± 0.02
VM	0.54 ± 0.06
EVM	0.54 ± 0.06

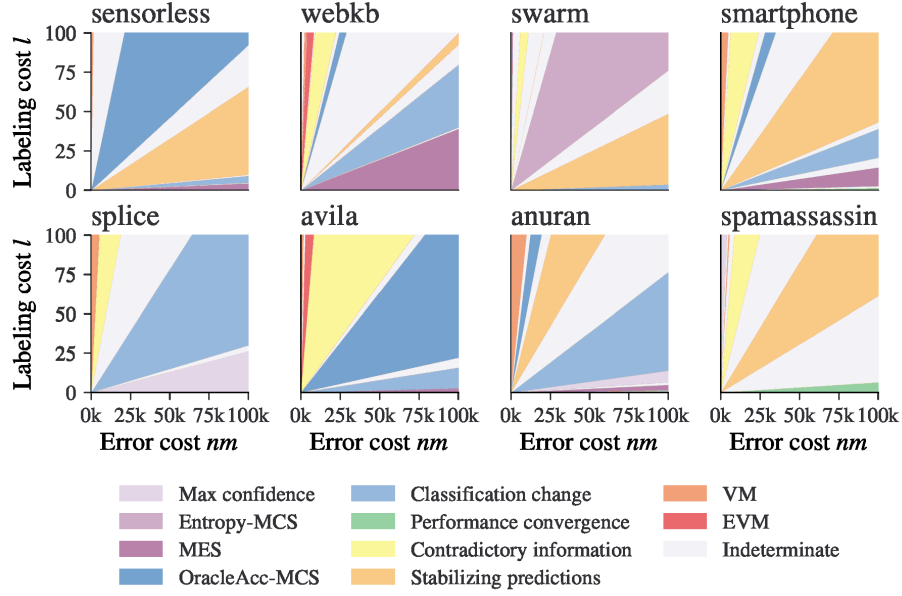


Figure 16: The regions in which each criterion is cost-optimal by dataset on an NN. Criteria that failed to stop were excluded. Light grey shows regions in which no SC was statistically dominant

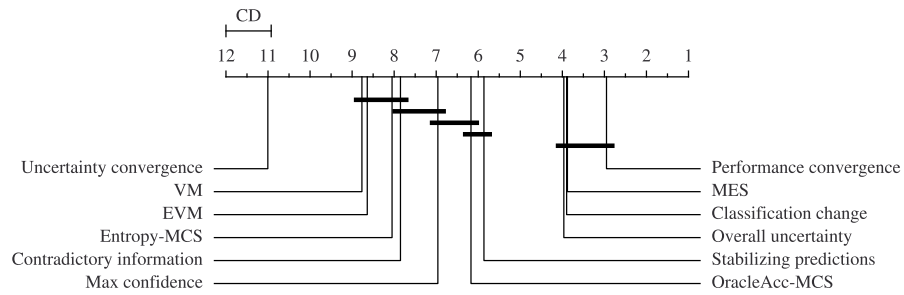


Figure 17: The relative cost of SC for an example scenario classifying mammo-grams using an NN

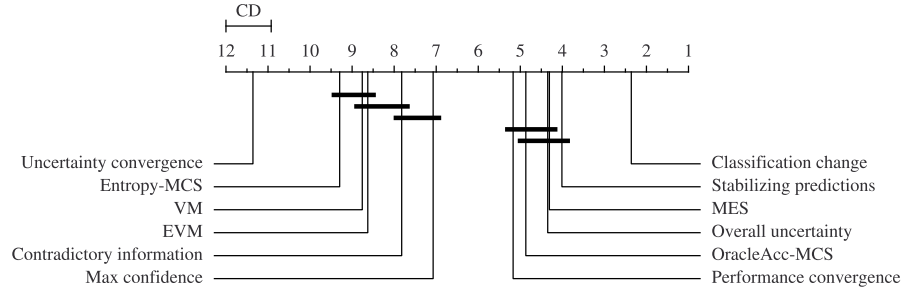


Figure 18: The relative cost of SC for an example scenario selecting potential customers using an NN

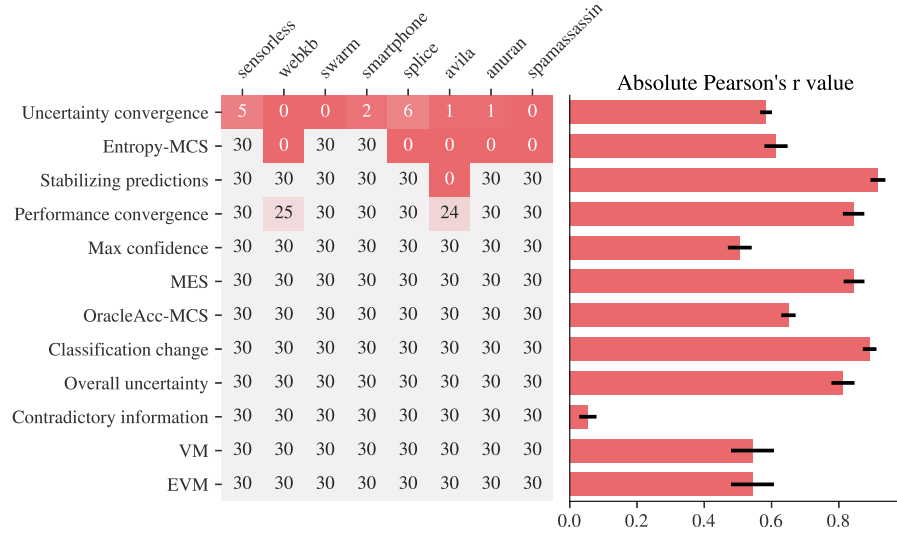


Figure 19: Left: the number of times each criterion stopped on each dataset using an NN out of 30 splits. Right: the Pearson correlation coefficient between each SC's metric and the classifier's accuracy. The black shows the standard error of the mean across datasets

5 SVM Class Imbalance

When applying active learning to imbalanced datasets SVMs can perform poorly. To demonstrate that the AL process proceeds normally we present a copy of Figure 2 from the paper except showing the area under the receiver operator characteristic (ROC AUC) instead of accuracy. This is shown in Fig. 20.

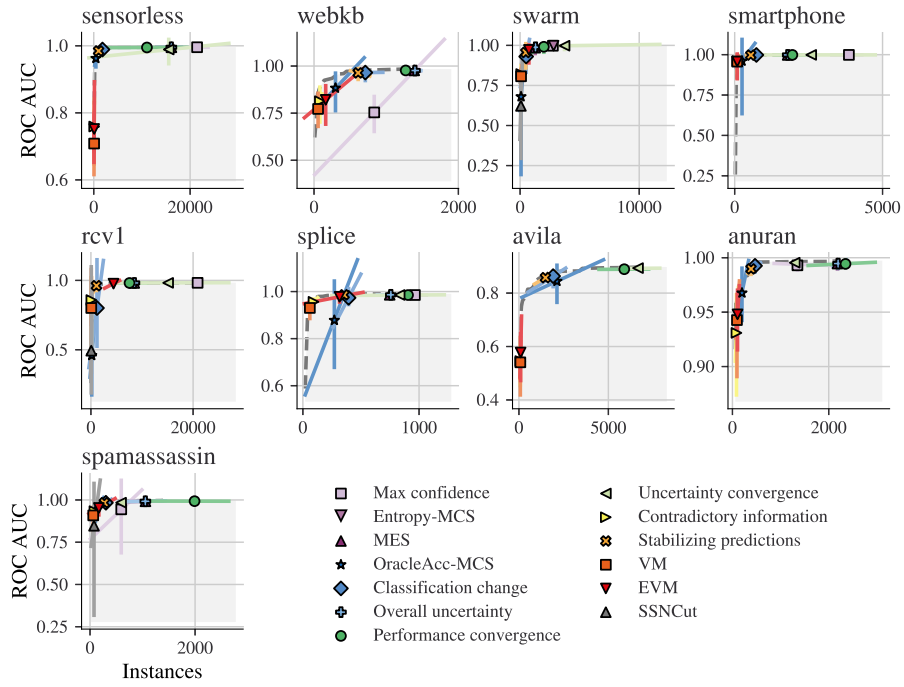


Figure 20: The ROC AUC and number of instances used by SC when evaluated for SVMs. The dashed grey line shows the Pareto frontier. Error bars were computed from 2.5% and 97.5% percentiles and transformed by PCA