

CSC730: Assignment 9

Huh, wonder what's out there?

Active Learning Based Rare Class Discovery

Kristophor Ray Jensen
Electrical Engineering and Computer Science
South Dakota School of Mines and Technology
Rapid City, United States
0009-0001-7344-349X

I. INTRODUCTION

The requirements for this assignment are as follows [1].

- 1) Get the provided datasets from D2L. Then for each dataset:
- 2) Visualize the data w/ labels using 2 or 3-D tSNE.
- 3) Write your own version of an active learning rare class discovery algorithm.
- 4) Run your code on the dataset and keep track of the number of classes discovered vs. number of queries.
- 5) Plot that (# classes discovered vs. # queries).
- 6) Rerun the same experiment using a random query strategy.
- 7) Plot the results from the random algorithm on the same plot.

II. METHODOLOGY

A. Active Learning Algorithm Overview

The main components of the active learning algorithm include the following [2]:

- **Dataset:** The dataset consists of labeled and unlabeled instances. The labeled instances are used to train the model, while the unlabeled instances are used to select instances for labeling.
- **Learning Algorithm:** The classifier is a machine learning model that is trained on the labeled data. The classifier is used to make predictions on the unlabeled data and select instances for labeling.
- **Query Strategy:** The query strategy determines which instances to query for labels. The query strategy selects instances that the model is most uncertain about, based on the current model's predictions.
- **Expert:** The expert is the user who provides labels for the selected instances. The expert labels the instances selected by the query strategy, and the labels are used to update the model.

B. Dataset

The MNIST data set consists of 70,000 labelled points. Each point is a 28x28 image of a digit from 0 to 9. The images are

reshaped to a 1D array of length 784 and normalized to the range [0, 1]. The data set is balanced, with an equal number of images for each digit class.

In our example we created a 20% split of the data set to be used as the test set. The training dataset was then put into a pool of unlabeled data.

A second list of the training data was created to be used as the labeled data. This method reminds of the pseudo-labeling method used in semi-supervised learning that we recreated in assignment 4.

Figure 1 shows a random selection of 16 nominal images from the training set.

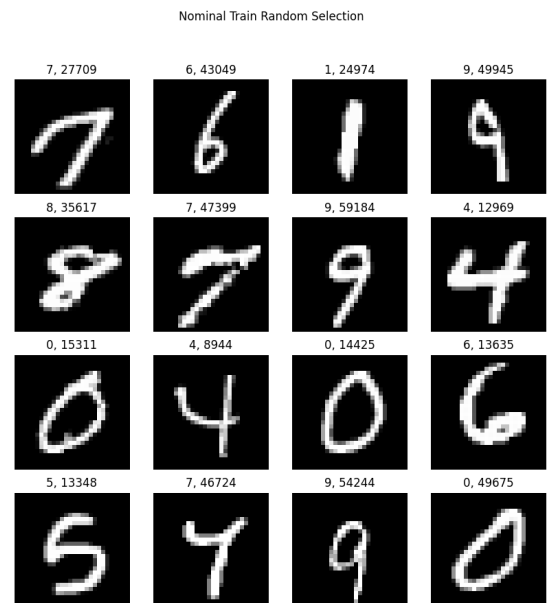


Figure 1. Random selection of nominal images from the training set

C. Learning Algorithm

There were two classifiers used in this assignment. The first classifier was a random forest classifier, and the second classifier was a support vector machine (SVM) classifier. The random forest classifier was used as the base classifier for the active learning algorithm. The SVM classifier was used to compare the performance of the active learning algorithm with a different classifier. The classifiers were trained on the labeled data and used to make predictions on the unlabeled data. The predictions were used to select instances for labeling based on the query strategy. The classifiers were updated with the newly labeled instances, and the process was repeated for either a maximum number of iterations or minimum single-class accuracy reaching a threshold of 95%.

D. Query strategy

In this assignment, we implemented the following query strategies:

- **Random Sampling:** Randomly select instances from the pool of unlabeled data for labeling. This strategy serves as a baseline for comparison with other query strategies.
- **Highest Confidence:** Select instances for labeling based on the classifier's highest confidence in its predictions. The highest confidence is calculated as the highest predicted probability for each instance.
- **Least Confidence:** Select instances for labeling based on the classifier's least confidence in its predictions. The least confidence is calculated as the difference between the highest and second-highest predicted probabilities for each instance.
- **Entropy Sampling:** Select instances for labeling based on the classifier's entropy of confidence in its predictions. The entropy is calculated as the negative sum of the predicted probabilities for each instance.
- **Lowest Vote:** Select instances for labeling based on the classifier's lowest vote in its predictions. The lowest vote is calculated as the lowest predicted probability for each instance.

E. Expert

The expert in this assignment was simulated by selecting instances from the pool of unlabeled data for labeling. The instances were selected based on the query strategy, and the labels were used to update the model. The expert was assumed to provide accurate labels for the selected instances.

F. Evaluation Metrics

Confusion matrices and accuracy scores were used to evaluate the performance of the active learning algorithm. The confusion matrix shows the number of true positives, false positives, true negatives, and false negatives for each class. The accuracy score is the proportion of correctly classified instances out of the total number of instances. The confusion matrix and accuracy score were calculated at each iteration of the active learning algorithm to track the performance of the model over time.

III. RESULTS AND DISCUSSION

There were a total of 10 training simulations ran for this assignment. These simulations were broken up by two classifiers and five query strategies. The classifiers used were the SVM and the Random Forest. The query strategies used were Random, highest uncertainty, highest confidence, entropy, and lowest vote. The results of these simulations are shown in the following tables.

Classifier	Highest Uncertainty	Highest Confidence	Random	Entropy	Lowest Vote
Rand Forest	232	557	339	256	203
SVM	175	632	270	855	195

Table I
NUMBER OF QUERIES REQUIRED TO REACH 95% ACCURACY

The training was performed to a stopping condition of either 95% accuracy for the worst performing class or 1000 iterations. None of the simulations reached the 1000 iteration stopping condition. The SVM classifier performed better than the Random Forest classifier in most cases.

IV. CONCLUSION

We were successful in implementing an active learning algorithm to train a classifier on a dataset of 60000 samples using only a small set of those samples. The algorithm was able to reach a 95% accuracy in far less than 1000 iterations.

The SVM classifier performed better than the Random Forest classifier in most cases. The highest uncertainty query strategy was the most effective query strategy for the SVM classifier. The lowest vote strategy was the most effective query strategy for the Random Forest classifier. The random query strategy proved to be a quite effective query strategy for both classifiers. The entropy query strategy provided mixed results for each classifier.

Due to the size of the graphs, they are added to the appendix.

REFERENCES

- [1] R. Loveland, "Assignment_9.pdf," From SDSMT D2L Website, 2024.
- [2] A. Tharwat and W. Schenck, "A survey on active learning: State-of-the-art, practical challenges and research directions," *Mathematics*, vol. 11, no. 4, p. 820, 2023. [Online]. Available: <https://doi.org/10.3390/math11040820>.

APPENDIX

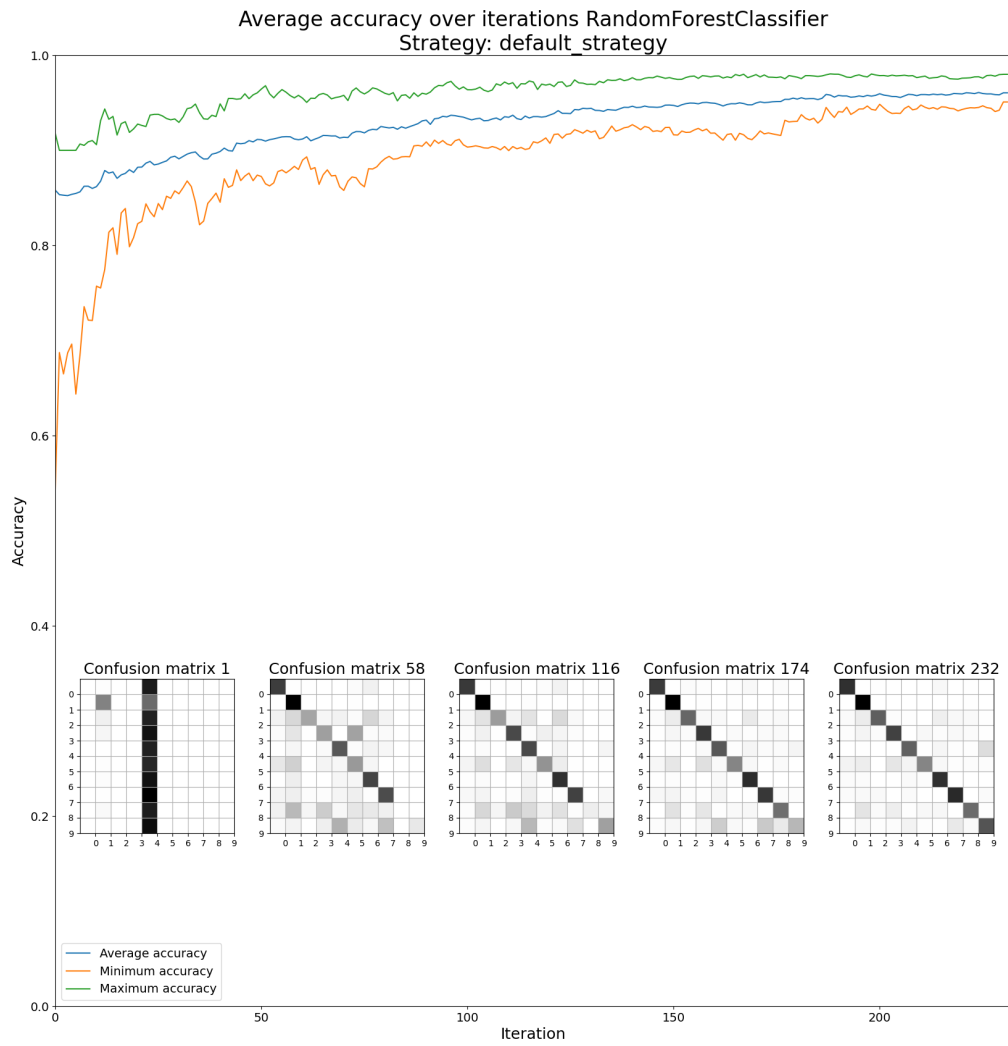


Figure 2. Random Forest Highest Uncertainty

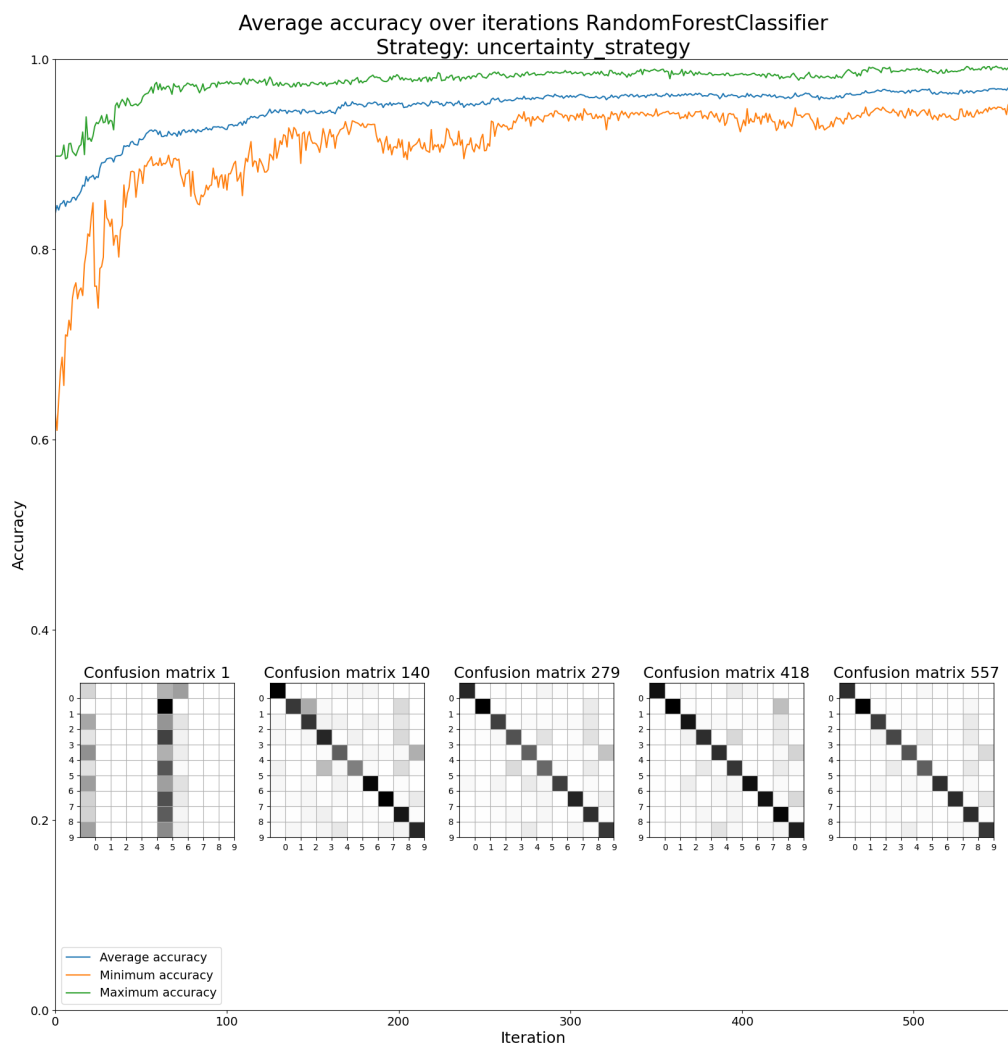


Figure 3. Random Forest Highest Confidence

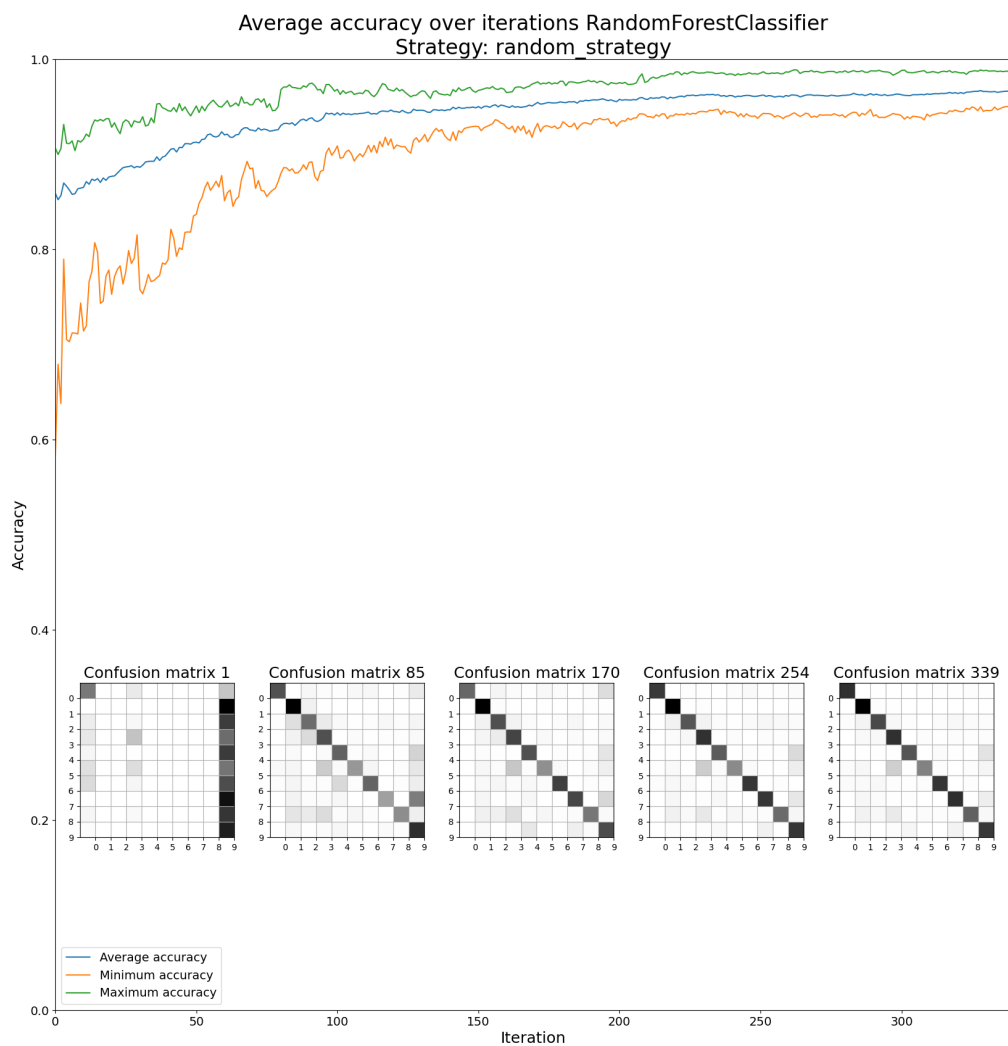


Figure 4. Random Forest Random

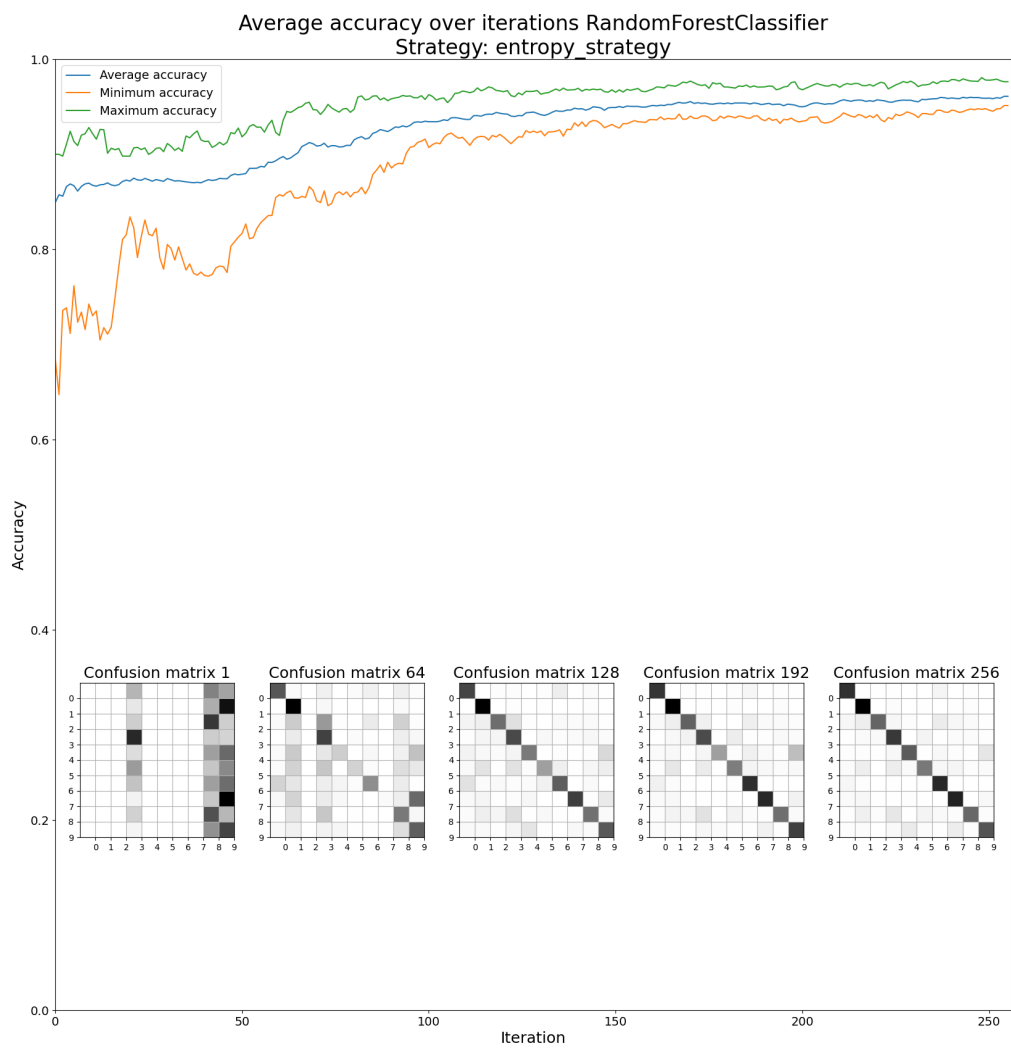


Figure 5. Random Forest Entropy

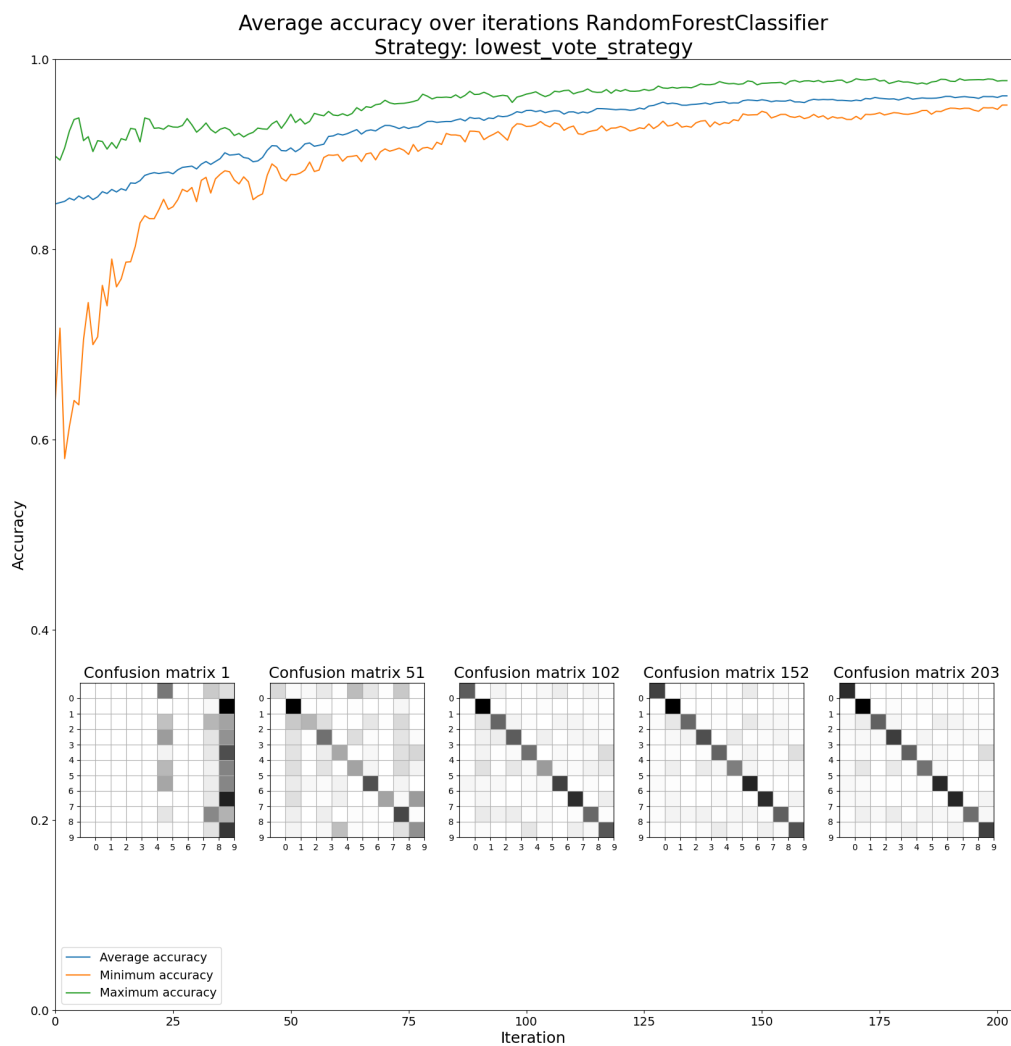


Figure 6. Random Forest Lowest Vote

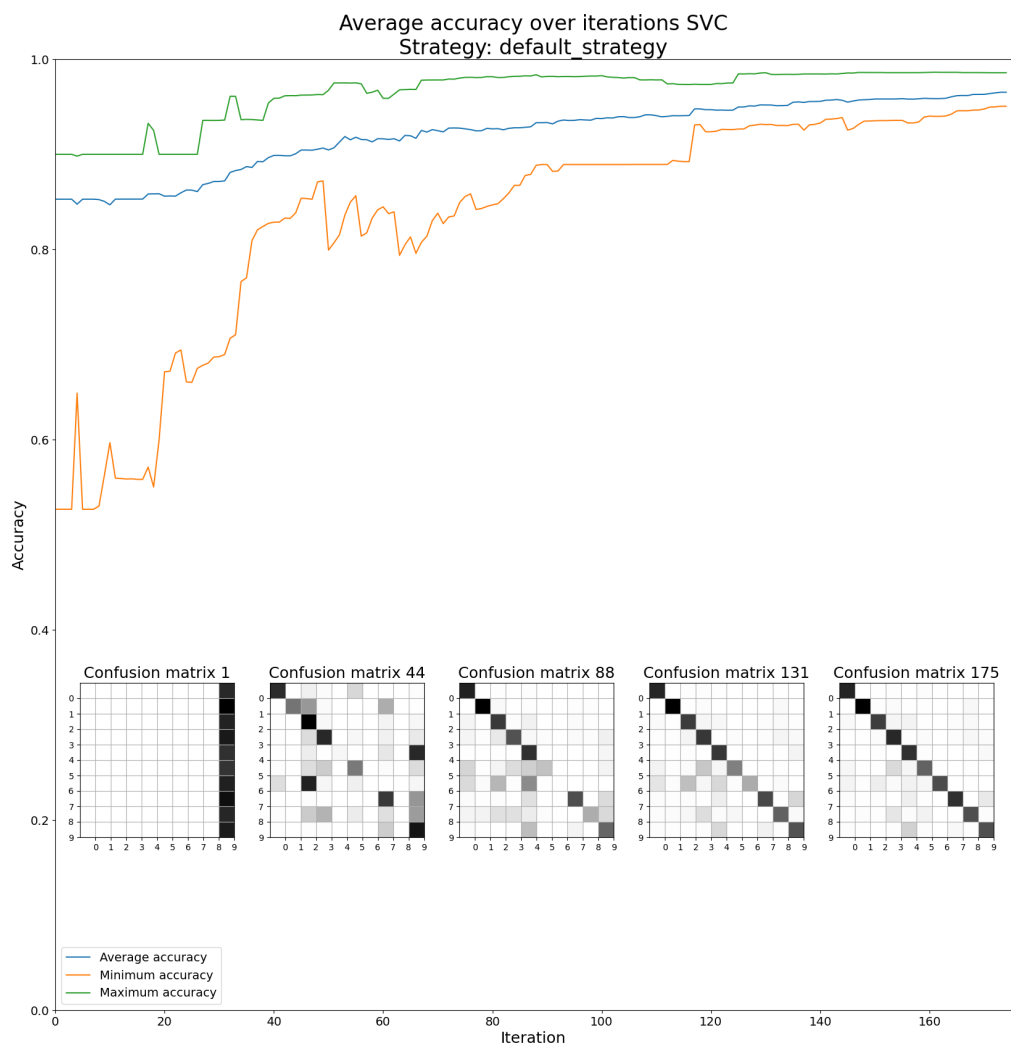


Figure 7. SVM Highest Uncertainty

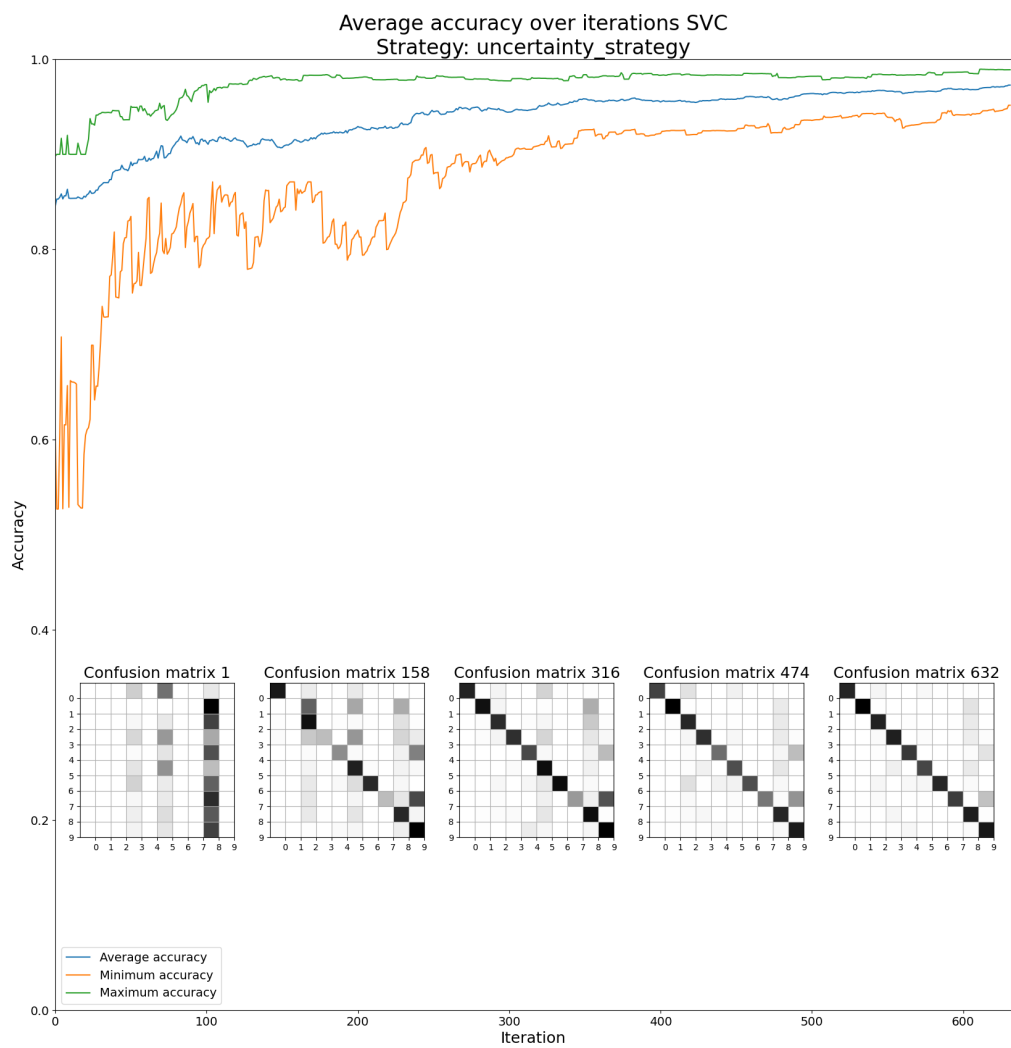


Figure 8. SVM Highest Confidence

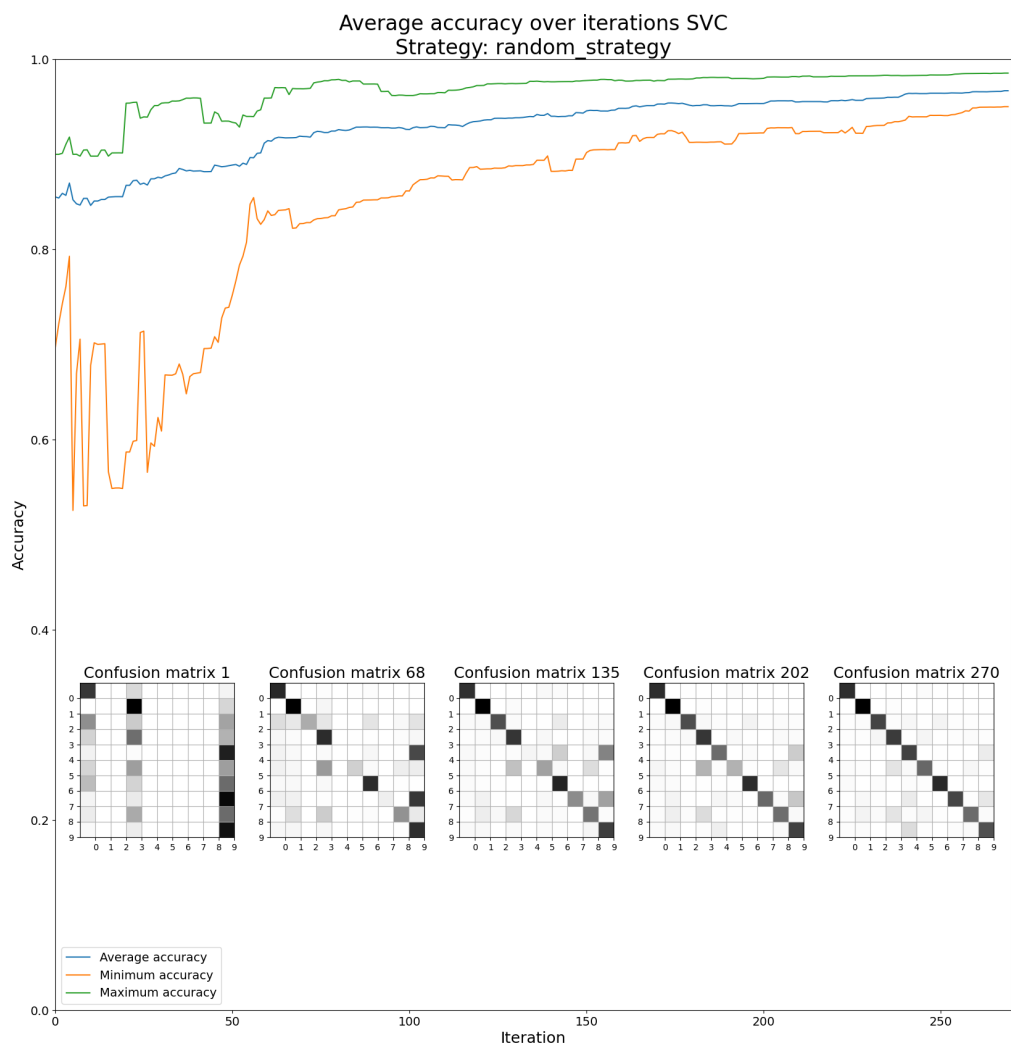


Figure 9. SVM Random

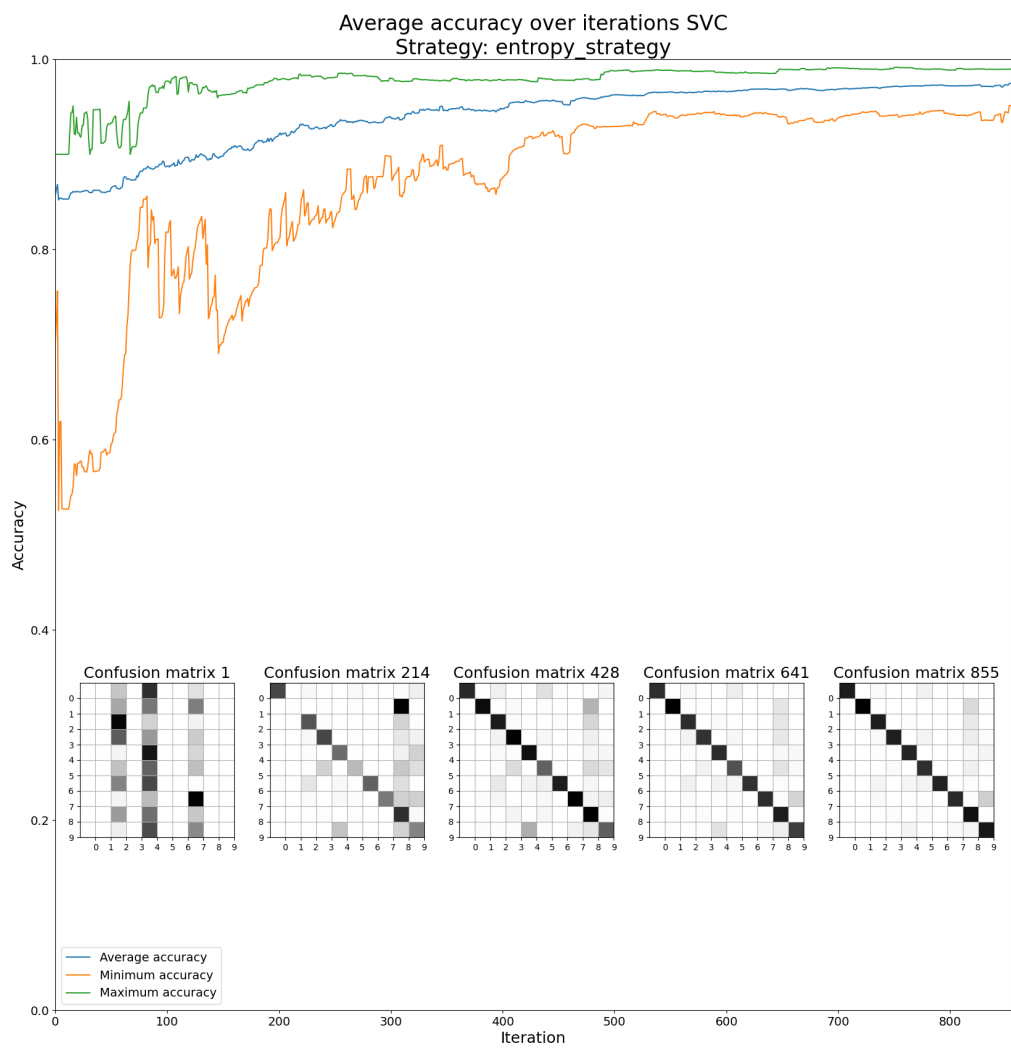


Figure 10. SVM Entropy

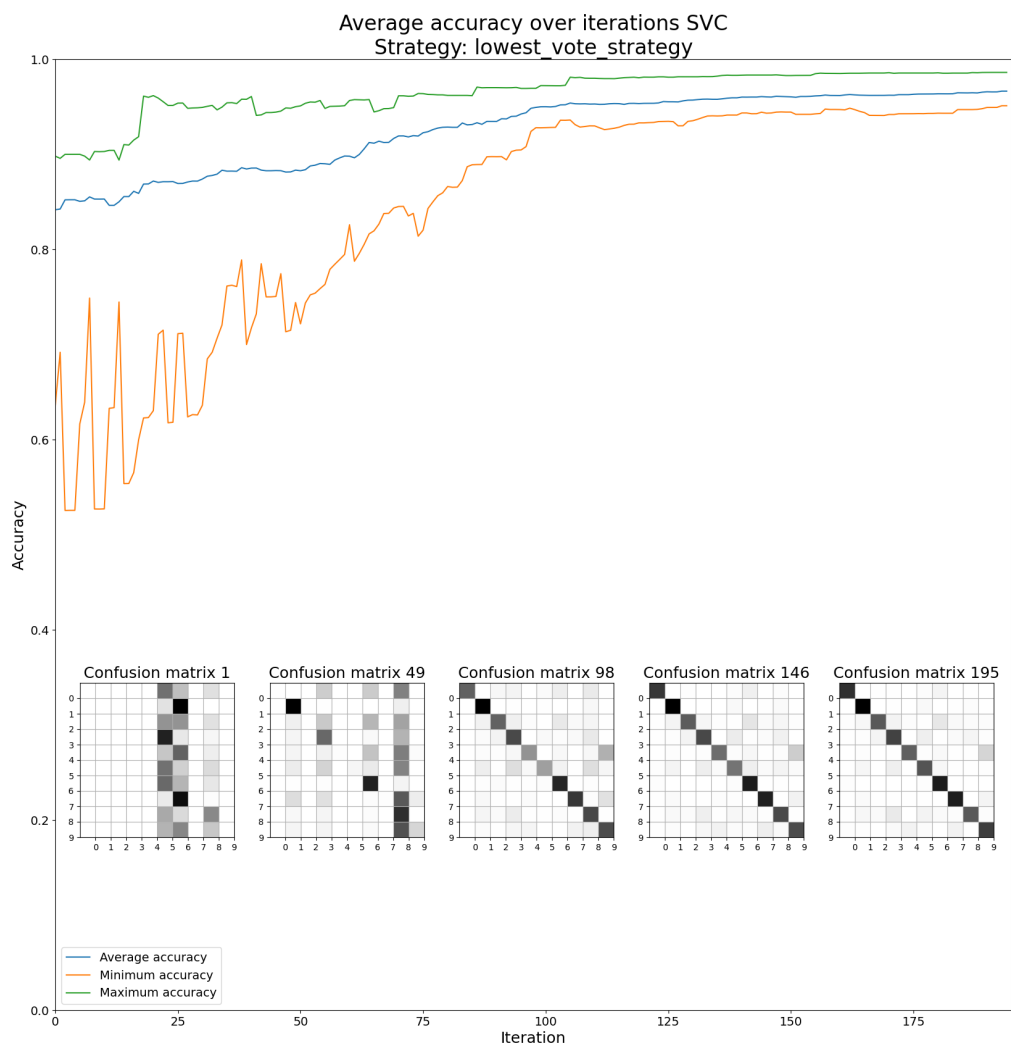


Figure 11. SVM Lowest Vote