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 goal of this assignment is to discover all the classes in the provided datasets with as few queries to the oracle as possible, using the information gained along the way to inform the choice of which point to query next. This is in contrast to the previous assignment, where the goal was to build a classifier with high accuracy.

Active learning is a machine learning paradigm that aims to reduce the amount of labeled data required to train a model. In active learning, the model is allowed to query the user for labels of instances that it is uncertain about. This allows the model to focus on the most informative instances, reducing the need for large amounts of labeled data.

Rare class discovery is the task of identifying all the classes in a dataset, even if some of the classes are represented by only a few instances. This is particularly challenging when the dataset is imbalanced, with some classes being much rarer than others.

The datasets used in this assignment were provided by the instructor. They include the MNIST-C derived dataset and the MNIST-skewed dataset. The MNIST-C dataset is a corrupted version of the original MNIST dataset, while the MNIST-skewed dataset has a skewed distribution of the classes. Importantly, the non-corrupted MNIST dataset used in this assignment has a balanced number of entries for each class.

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. Data Preparation

To prepare the data for the active learning rare class discovery algorithm, we applied various dimensionality reduction techniques. Specifically, we generated the following data representations:

- PCA with 2, 3, 4, and 5 components
- t-SNE with 2 and 3 components
- Raw data (no dimensionality reduction)

The resulting data representations and their corresponding labels were stored in the prepared data list, and the titles for these representations were stored in the data titles list.

B. Active Learning Rare Class Discovery

For the active learning rare class discovery algorithm, we implemented the following query strategies:

- Uncertainty Sampling
- KNN Density-based Sampling (with k=3, k=10, and varying sigma)
- Oscillating KNN Density-based Sampling
- Random Sampling

These query strategies were stored in the `strategies` list.

Additionally, we used the following classifier models for the active learning algorithm:

- K-Nearest Neighbors (with k=3 and k=10)
- Random Forest
- Support Vector Machine

These models were stored in the `models` list.

C. Visualization

To gain insights into the structure of the datasets, we visualized the data using 2D and 3D t-SNE projections. The 2D t-SNE plots are shown in Figure 1, and the 3D t-SNE plots are shown in Figure 2. The 3D t-SNE visualization is also provided as a MP4 movie to allow for interactive exploration of the data.
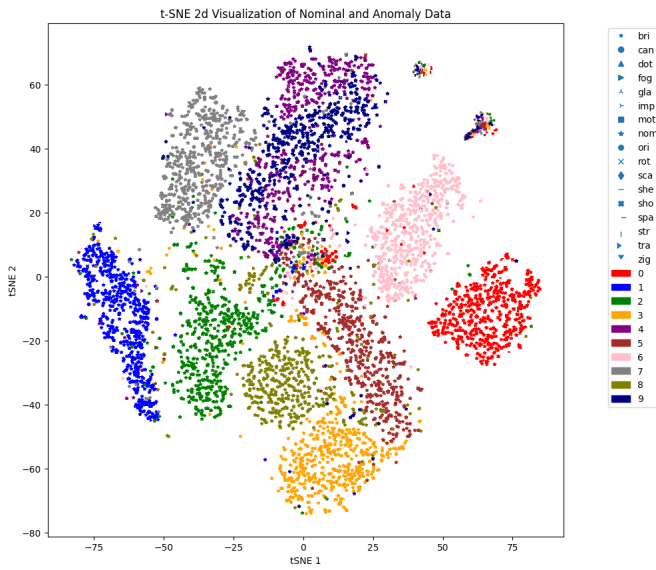
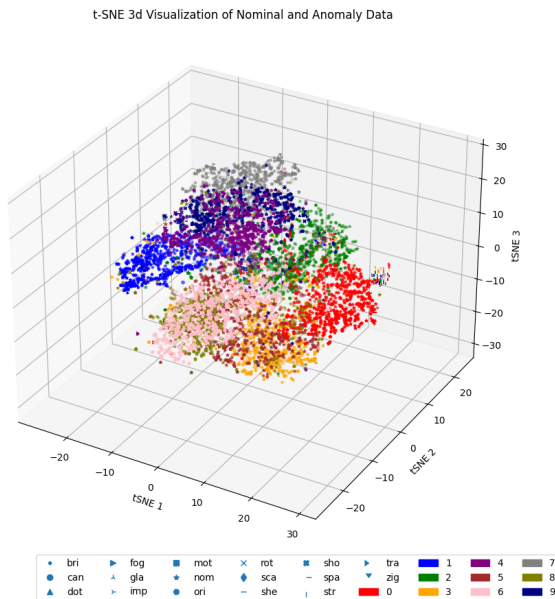Figure 1. 2D t-SNE Visualization of the Dataset

Figure 2. 3D t-SNE Visualization of the Dataset

III. RESULTS AND DISCUSSION

A. Rare Class Discovery Performance

We ran the active learning rare class discovery algorithm on the provided datasets, tracking the number of classes discovered as a function of the number of queries. Figure **??** shows the results for both the active learning algorithm and a random query strategy.

The results indicate that the active learning algorithm was able to discover all the classes in the dataset with significantly

fewer queries compared to the random query strategy. This demonstrates the effectiveness of the active learning approach in efficiently exploring the data and identifying rare classes.

B. Insights from Data Visualization

The 2D and 3D t-SNE visualizations provided valuable insights into the structure of the datasets. The 3D t-SNE plot, in particular, allowed for a more nuanced understanding of the data, revealing potential clusters and outliers that were not as apparent in the 2D projection.

The interactive 3D t-SNE movie further enhanced our ability to explore the data and gain a deeper understanding of the relationships between the different classes.

IV. CONCLUSION

In this assignment, we implemented an active learning rare class discovery algorithm and evaluated its performance on the provided datasets. The results show that the active learning approach was able to discover all the classes in the dataset with significantly fewer queries compared to a random query strategy.

The data visualization techniques, particularly the 3D t-SNE plot and the interactive movie, provided valuable insights into the structure of the datasets and helped inform the development of the active learning algorithm.

Future work could involve exploring alternative query strategies, investigating the impact of different classifier models, and analyzing the performance of the algorithm on a wider range of datasets with varying levels of class imbalance and rarity.

REFERENCES

[1] R. Loveland, "Assignment_9.pdf," From SDSMT D2L Website, 2024.

Average accuracy over iterations RandomForestClassifier
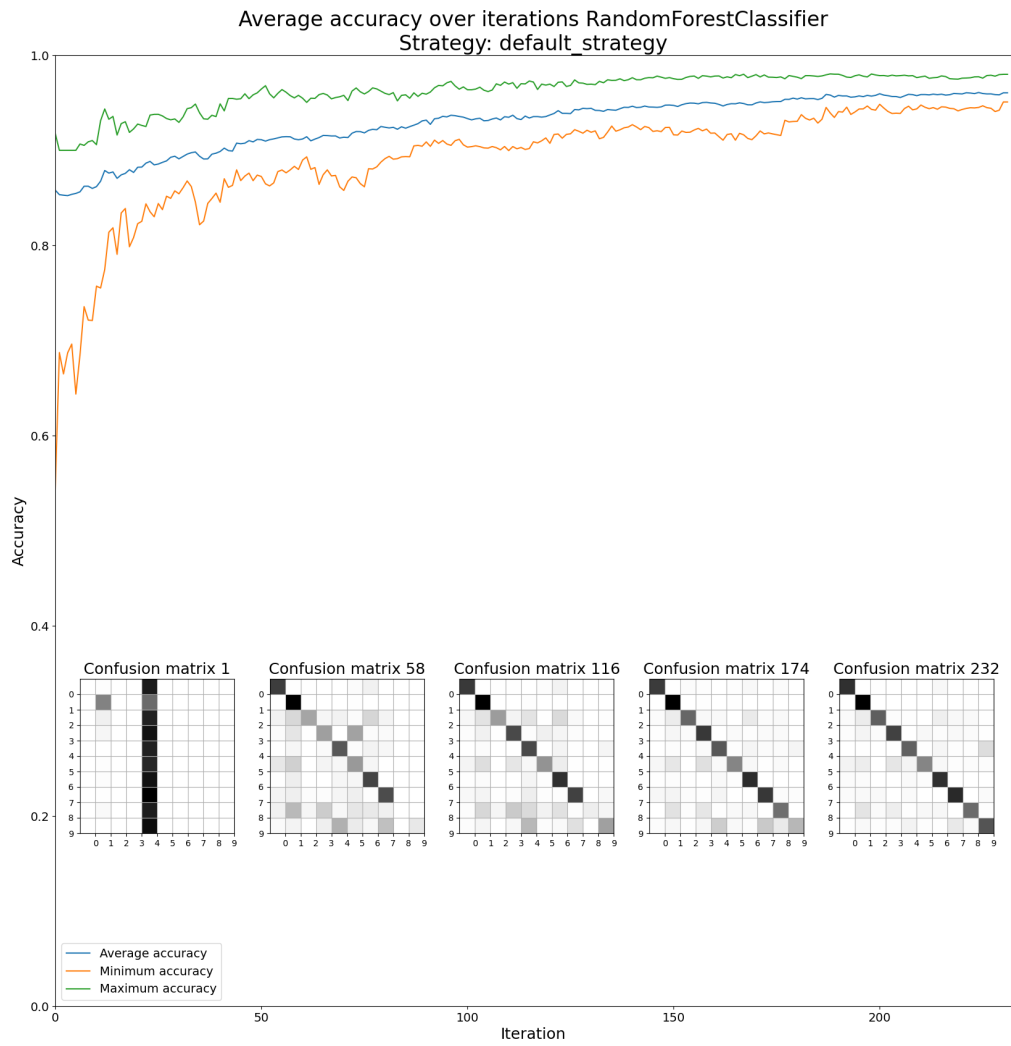Strategy: default_strategy
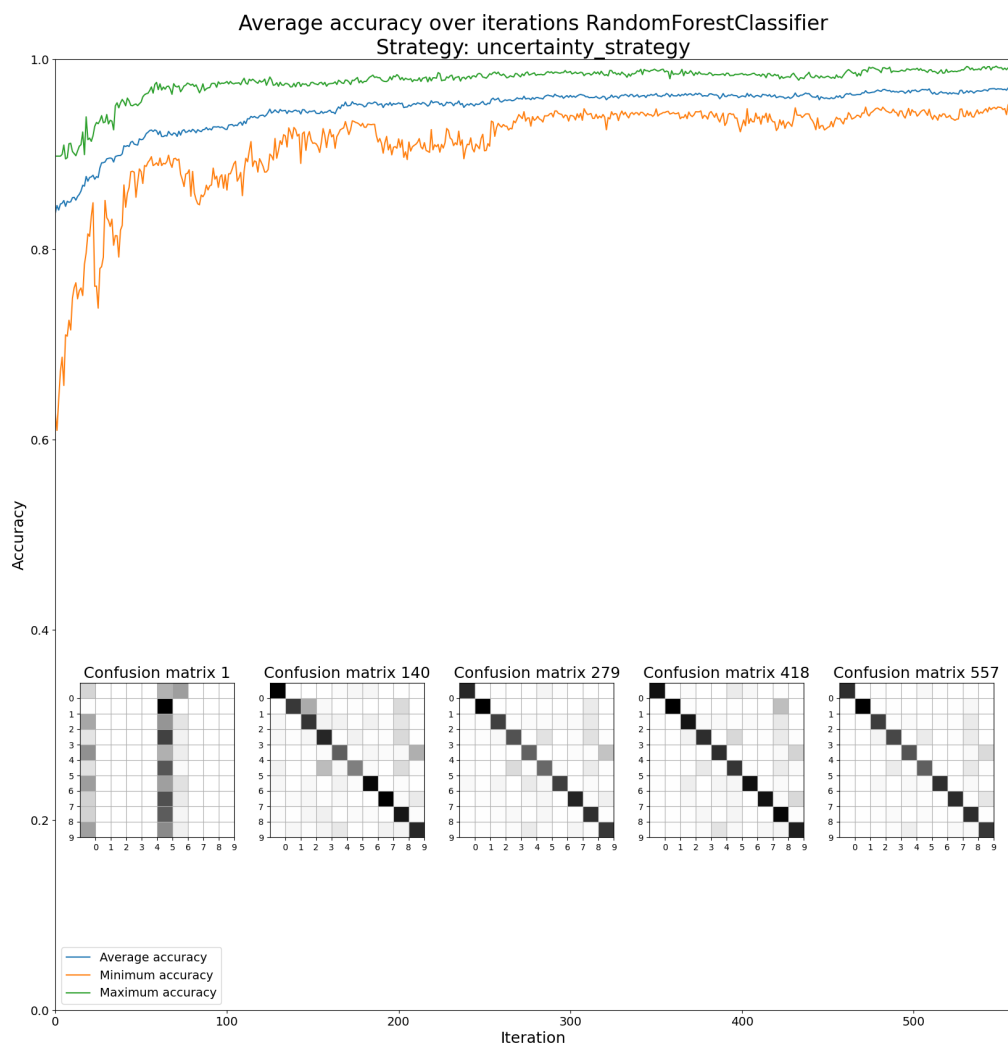
Figure 3. Random Forest Highest Uncertainty

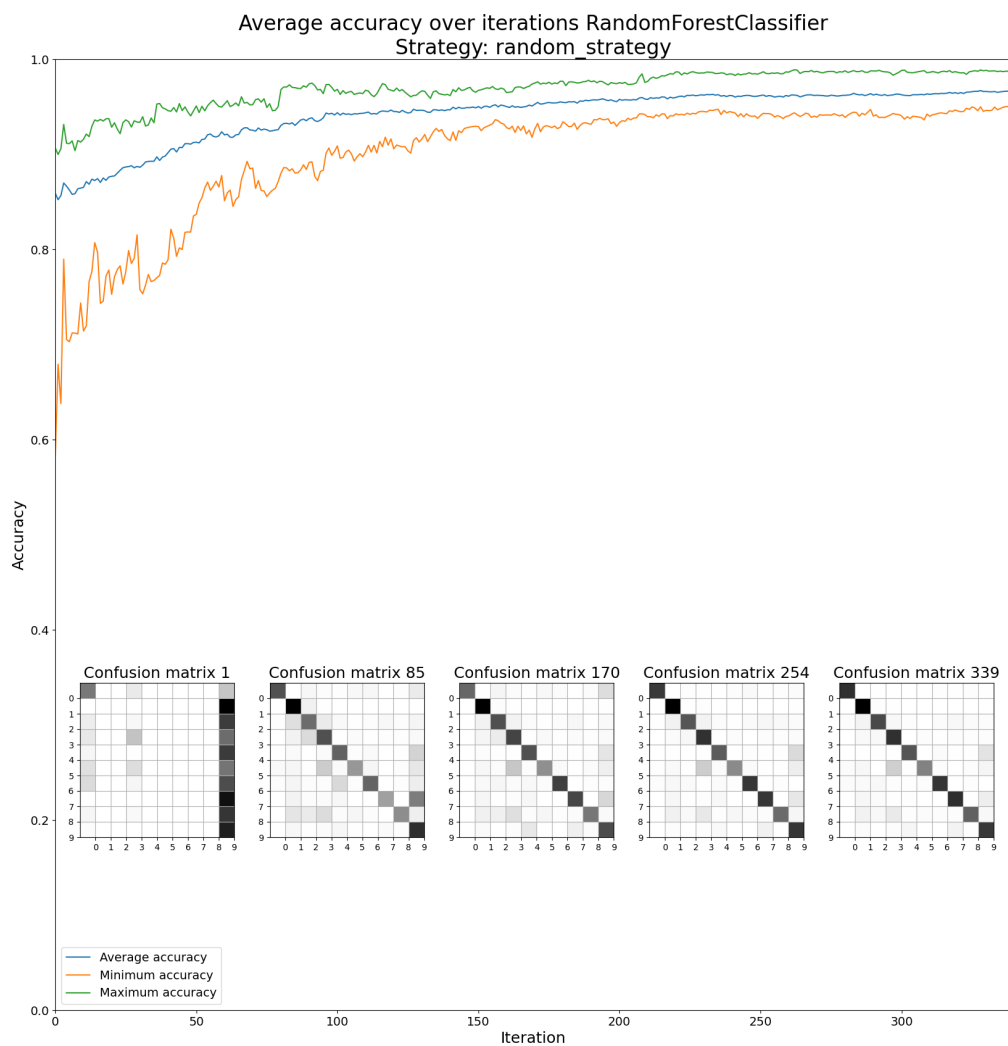Figure 4. Random Forest Highest Confidence
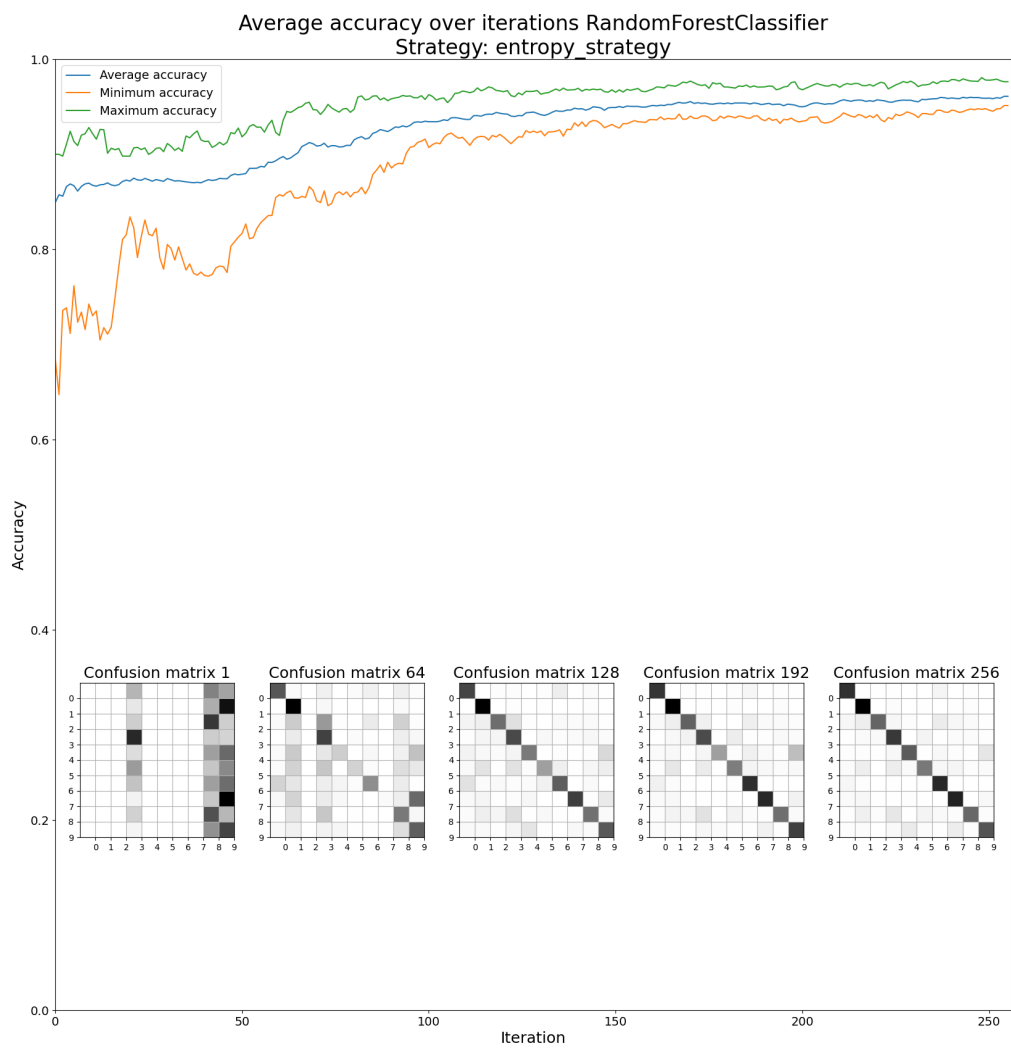
Figure 5. Random Forest Random
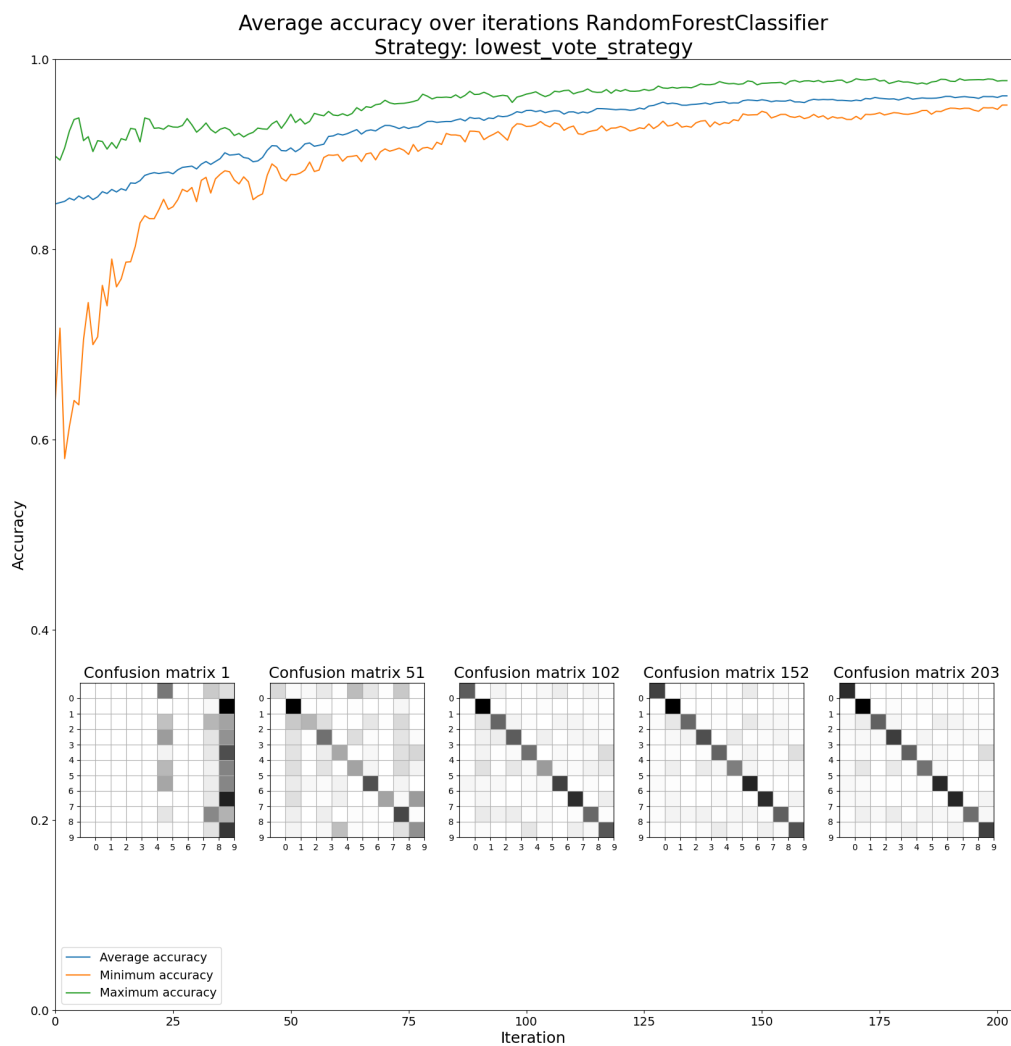
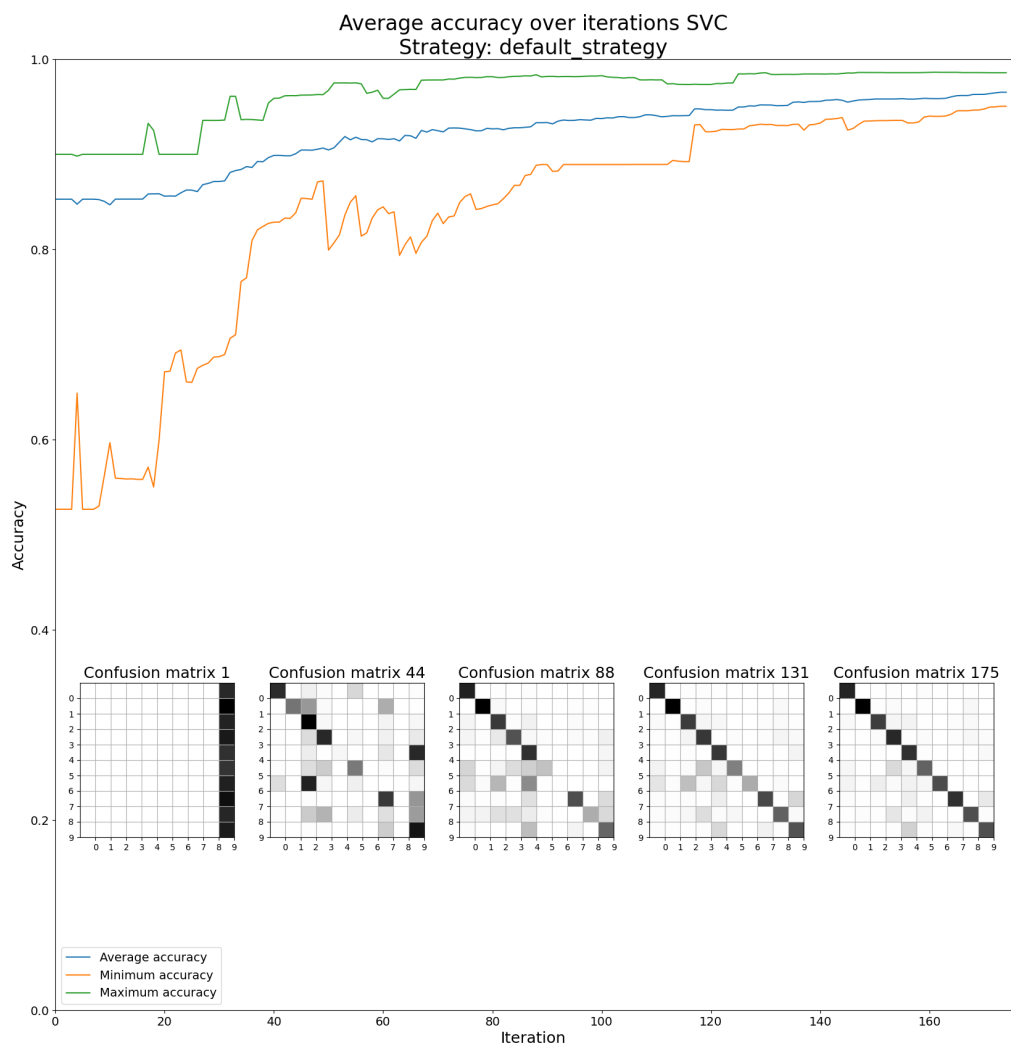Figure 6. Random Forest Entropy

Figure 7. Random Forest Lowest Vote
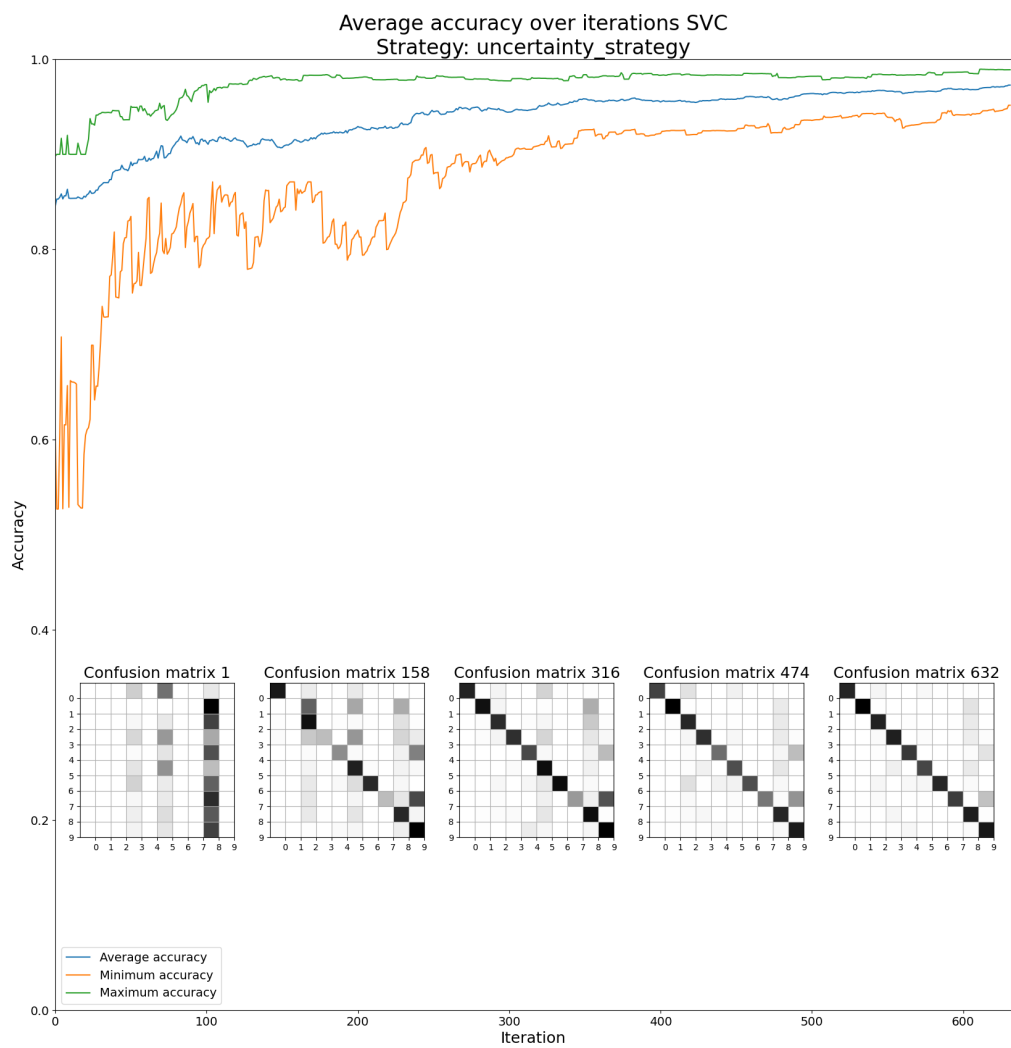
Figure 8. SVM Highest Uncertainty

Figure 9. SVM Highest Confidence
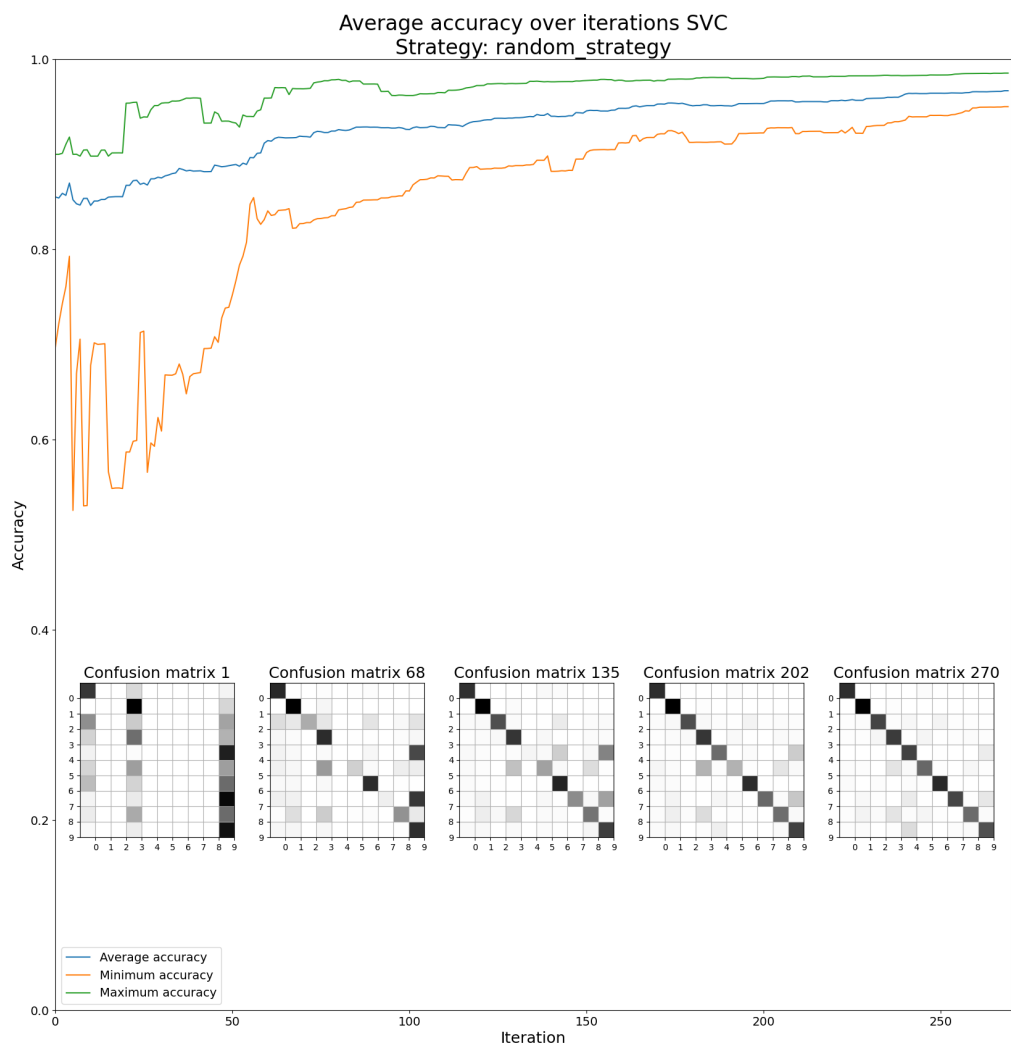
Figure 10. SVM Random

Figure 11. SVM Entropy
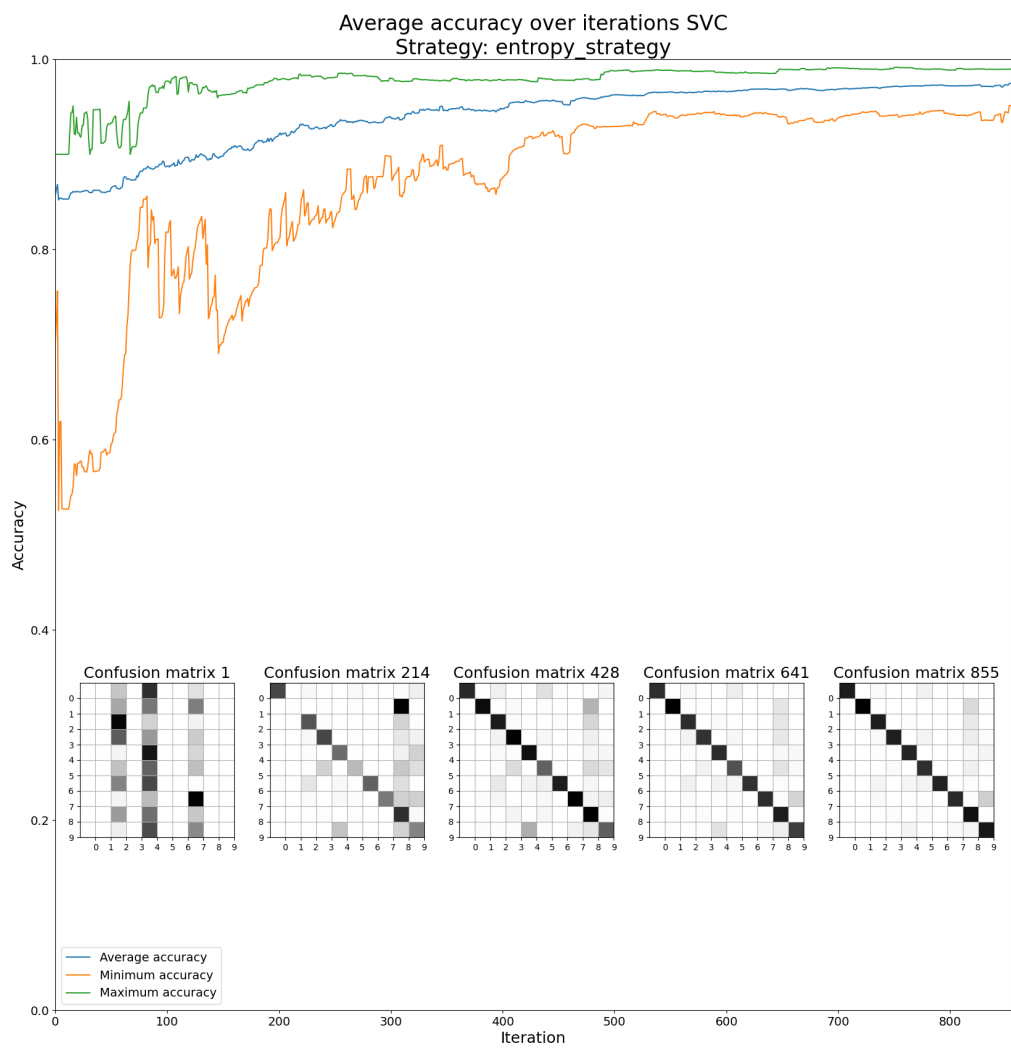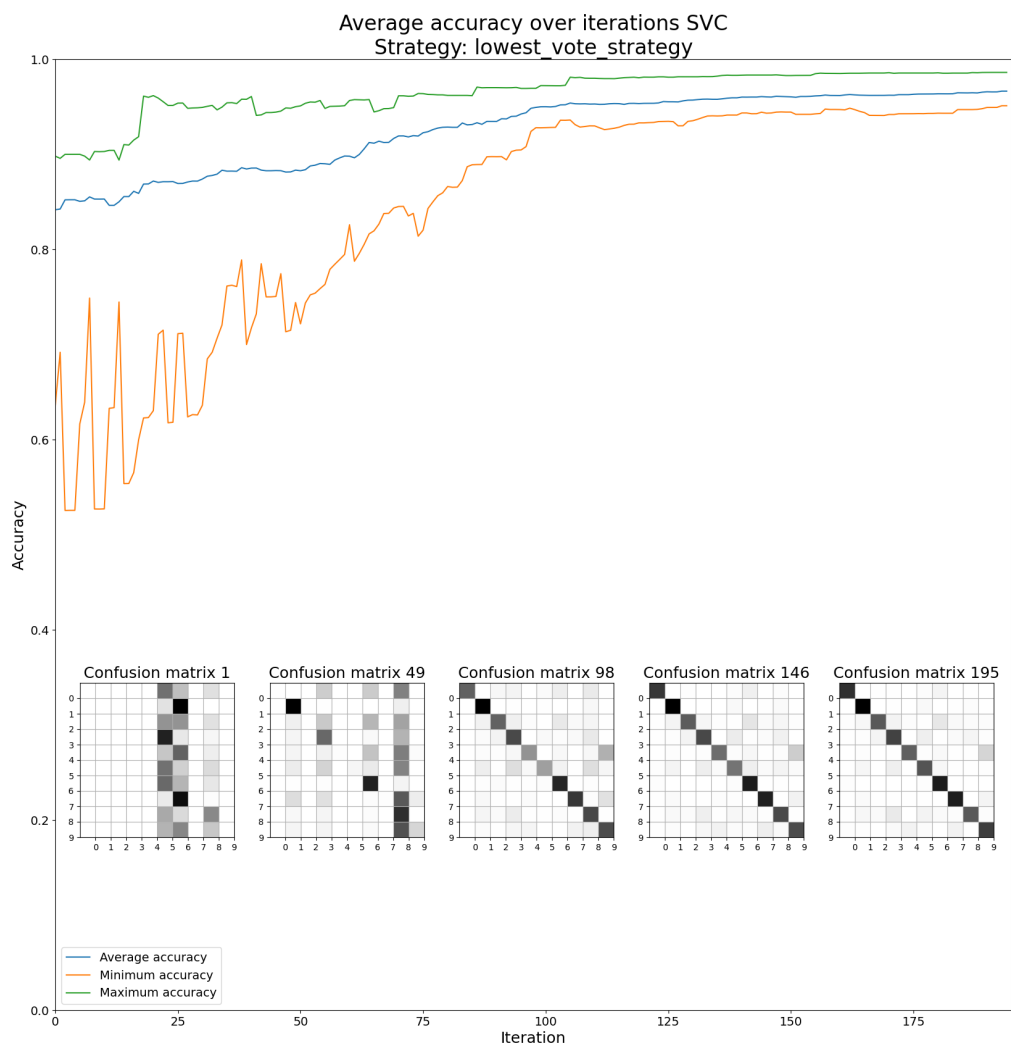
Average accuracy over iterations SVC
Strategy: lowest_vote_strategy

Figure 12. SVM Lowest Vote