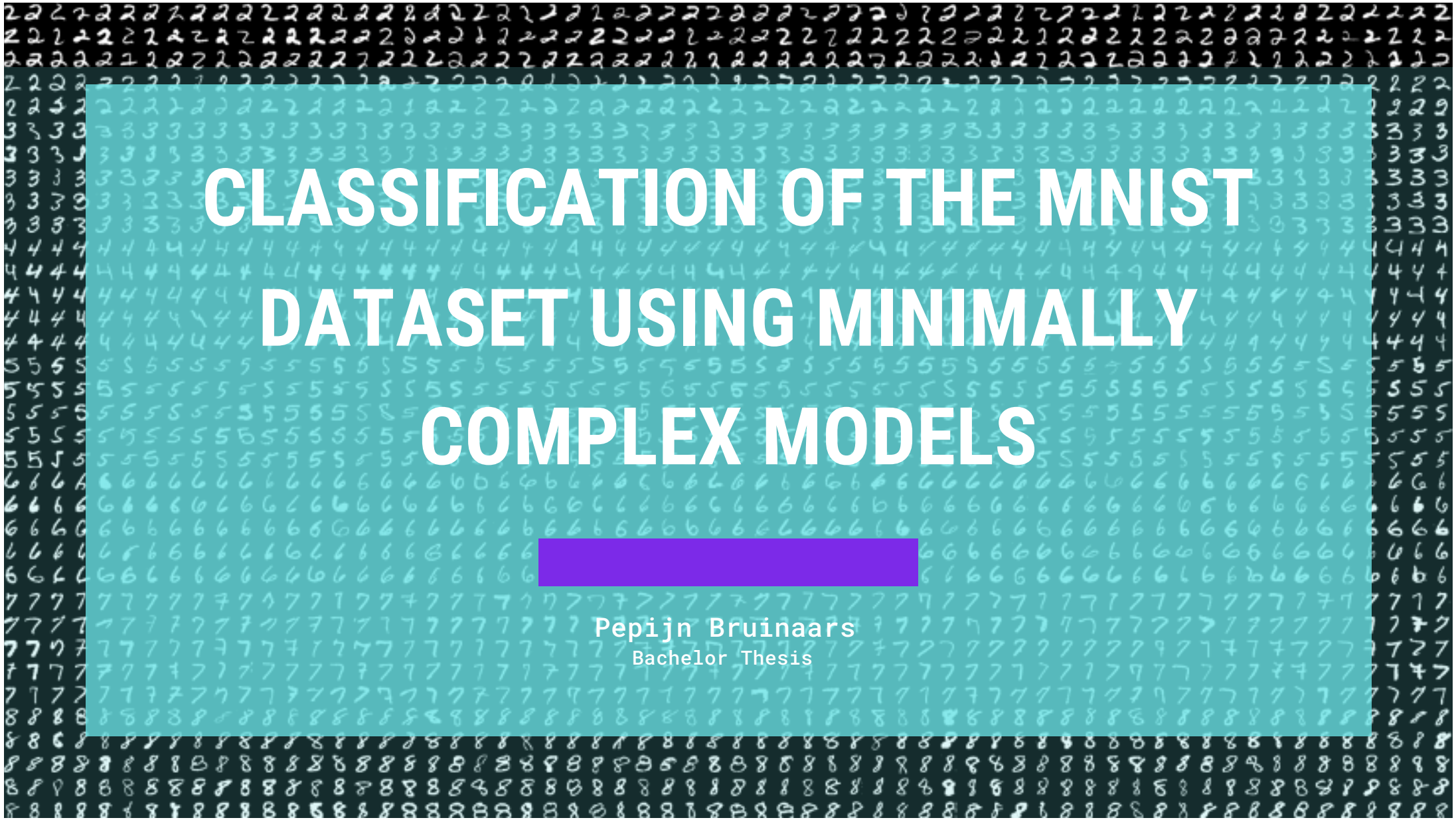


CLASSIFICATION OF THE MNIST DATASET USING MINIMALLY COMPLEX MODELS

Pepijn Bruinaars

Bachelor Thesis

RESEARCH QUESTION

ARE MINIMALLY COMPLEX MODELS SUITABLE FOR CLASSIFICATION TASKS AND WHAT IS THEIR PERFORMANCE COMPARED TO OTHER CLASSIFIERS?

MINIMALLY COMPLEX MODELS

SPIN MODELS

- ICCs
- Better explainability

CLASSIFICATON

- No parameter fitting
- Goodness-of-fit vs. generalizability
- Less data

NAIVE BAYES CLASSIFIERS

BAYES RULE

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)}$$

$$P(C_k|x) = \frac{P(x|C_k)P(C_k)}{P(x)} \propto P(x|C_k)P(C_k)$$

CLASSIFICATION

$$C = \{P(C_0|x), P(C_1|x), \cdots, P(C_n|x)\}$$

$$f(x) = \hat{y}(x) = \underset{c}{\operatorname{argmax}}\, P(C_k|x)$$

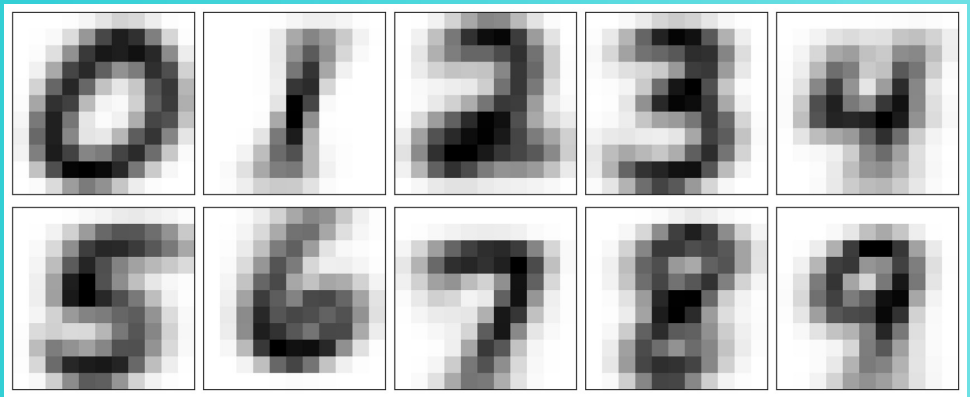
METHOD

1. PRE-PROCESSING

2. CREATING THE CLASSIFIER

3. EVALUATION

PRE-PROCESSING



‘Average’ digits of the compressed data set

COARSE-GRAINING

- Variable limit (128)
- Coarse-graining using gray values
- 2x2 squares → single pixel

BALANCING

- Highly unbalanced data set
- Possibly creates biases
- Created uniform data set

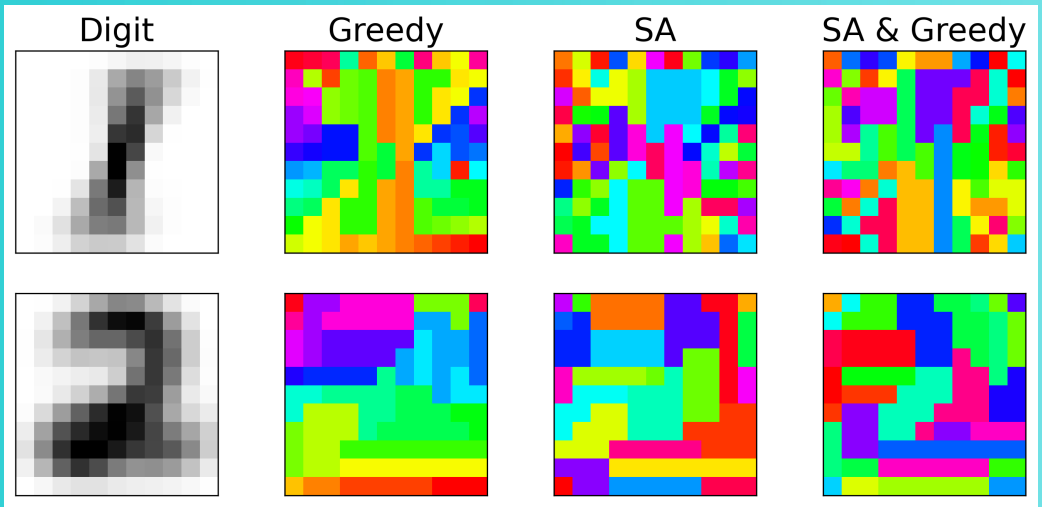
SELECTING MCMs

- Greedy Search
- Simulated Annealing

CLASSIFICATION

- Select MCMs for each digit
- Calculate $P(C_k|x)$ for each MCM
- 'Predict' C_k with the largest $P(C_k|x)$

CREATING THE CLASSIFIER



Community assignments for the digits 1 and 2 for various selection methods

CREATING THE CLASSIFIER

LIMITATIONS OF MCM-CLASSIFIER

- Reject option
- Only low variable counts

RBM-BASED CLASSIFIER

- Digit features using RBM
- Classification using logistic regression

EVALUATION

CLASSIFICATION METRICS

- Accuracy
- Non-rejected accuracy
- Precision
- Recall
- F1-score
- Fitting/selection time (s)

HYPER-PARAMETERS

- Various training set sizes
- 5 different combinations of selection parameters

RESULTS

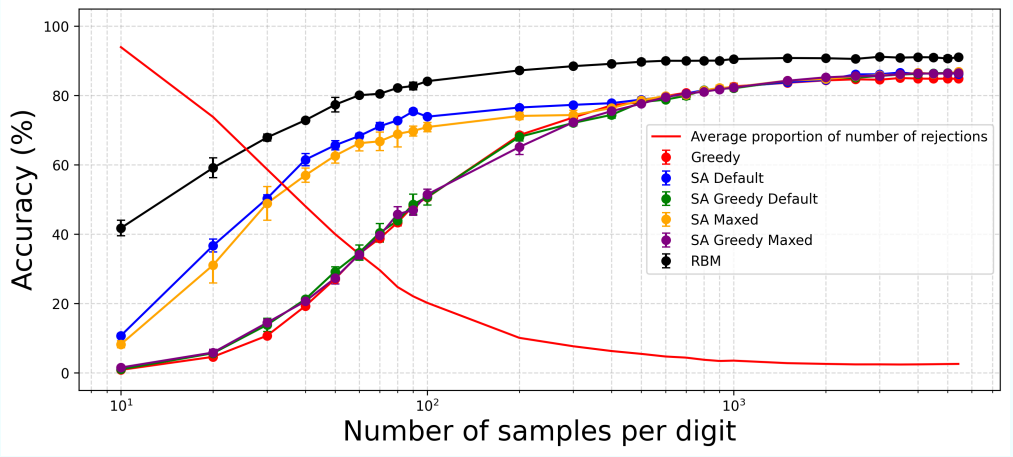
ACCURACY

- Relatively good
- RBM is better

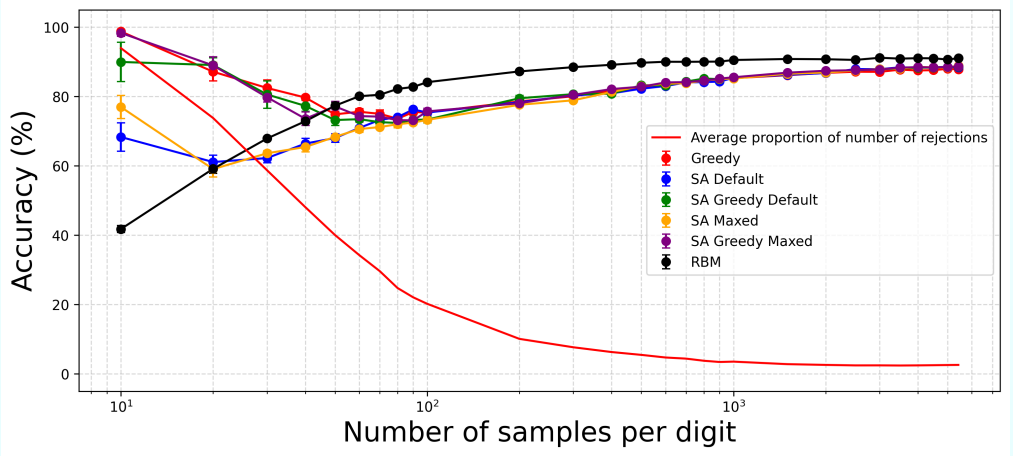
NON-REJECTED ACCURACY

- Extremely high for low data set sizes
- Indicates high uncertainty

(True) accuracy of the different classifiers



Non-rejected accuracy of the different classifiers



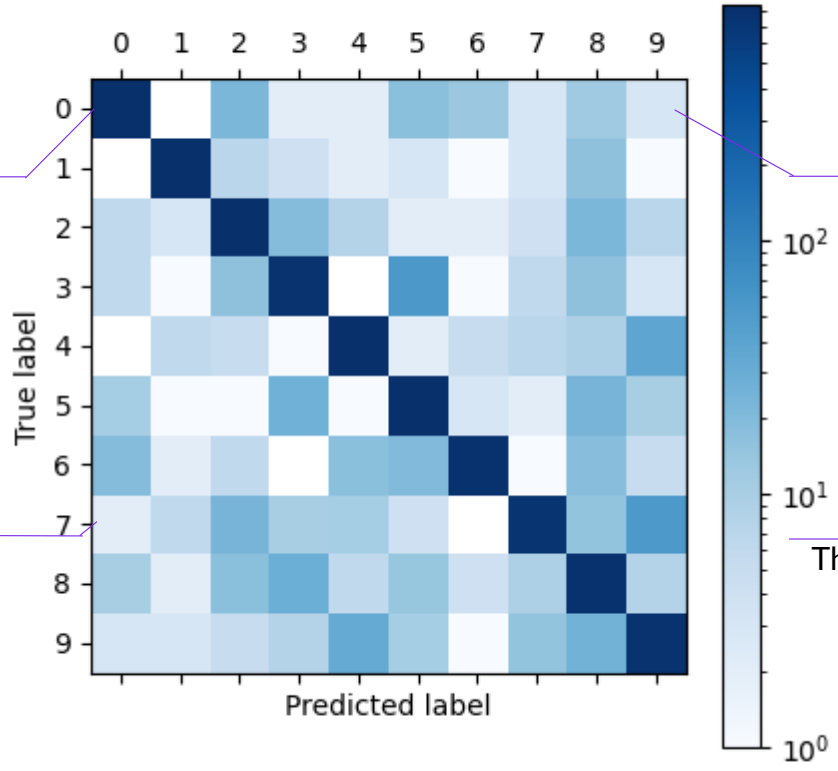
CONFUSION MATRIX

ACCURACY

How many of the predictions were correct?

PRECISION

How many of the samples classified as some digit were truly that digit?



RECALL

How many of the samples labeled as some digit were identified as such?

F1-SCORE

The F1-score is the harmonic mean of the recall and precision.

RESULTS

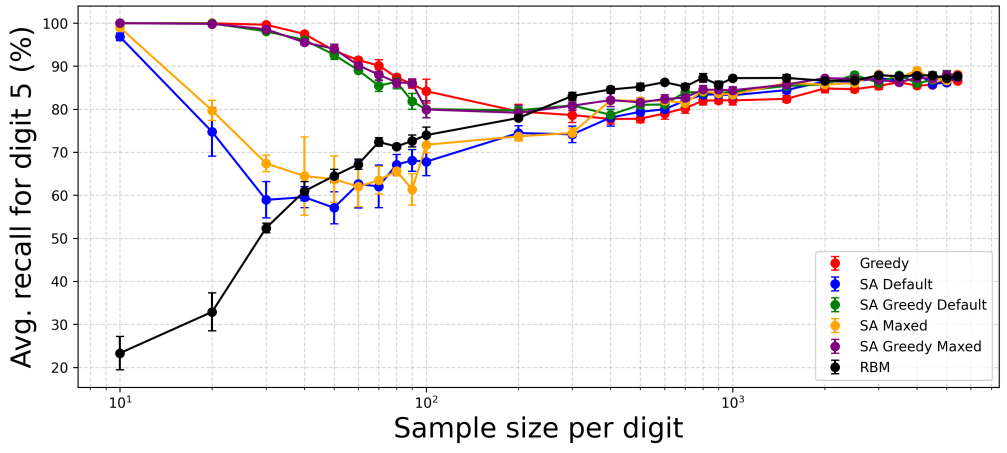
RECALL

- MCM-classifiers comparable as data set grows
- Only includes non-rejected samples

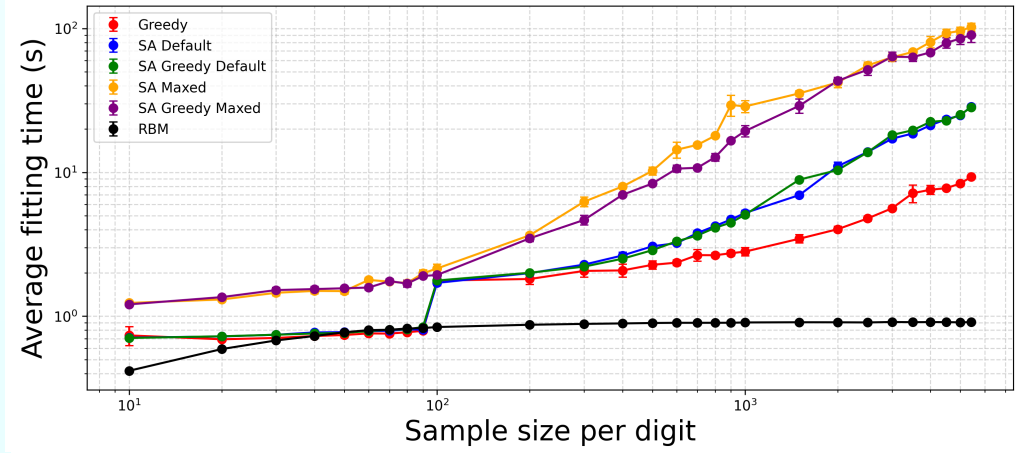
FITTING TIME

- RBM faster than any MCM-classifier
- Jump in fitting time at 100 samples
- Greedy fastest of MCM-classifiers

Recall of the classifiers for digit 5



Fitting/selection times of the classifiers



DISCUSSION

REJECTION OPTION

- Still available information
- Flipping some pixels in rejected images
- Basis change

COMPARISON TO RBMs

- Not a ‘fair’ comparison
- More optimized

LIMITATIONS

- 128 variables
- GPU Programming (CUDA)

CONCLUSION

LIMITATIONS OF MCM-CLASSIFIER

- Reject option
- Only low variable counts

TAKEAWAYS

- Show good promise
- Potentially improved explainability

THANKS FOR LISTENING