Lehrstuhl für Bildverarbeitung
Institute of Imaging & Computer Vision
RWTH AACHEN
RHEINISCH-WESTFÄLISCHE TECHNISCHE HOCHSCHULE AACHEN

Predicting the Benefit of Sample Size Extension in Multiclass k-NN Classification

Christian Kier[1] and Til Aach[2]

[1]Institute for Signal Processing, University of Lübeck, D-23538 Lübeck, kier@isip.uni-luebeck.de
[2]Institute of Imaging and Computer Vision, RWTH Aachen University, D-52074 Aachen

Introduction

- Obtaining training data is very costly in industrial classification problems.
- Classifier quality crucial for economic success.
- Usually multiclass classification problems.
- When do we have enough samples?

Objectives

- Give hint on best possible classifier performance with given problem.
- Gather enough training samples for desired classifier quality.
- Avoid gathering of too many samples.
- Make concrete statement on training set extension process.

Assumptions

- Asymptotical error rate $e_\infty = \lim_{n \to \infty} e(n)$ exists for chosen classifier.
- Error rate $e(n)$ converges towards e_∞.
- Specific problem is given.
- 3-NN is used for classification.

Idea

- Model $e(n)$ through measurements on data set of size N.
- Decrease data set size by randomly removing samples.
- For every N estimate $e(N)$ by cross-validation.
- Fit model function $e_m(n)$ to measurements to extrapolate $e(n)$ beyond N.
- Derive $e_m(n)$ based on error probability. For details see paper. Use k-NN probability density estimates given by [5]:

$$\overline{p(x|\omega_i)} = \frac{k}{n_i \cdot V_k} = \frac{k}{n_i \cdot c_d \cdot r_k^i(x)^d}$$

  with n_i being number of samples of class ω_i and V_k volume of the hypersphere around x spanning over k neighbours.
- Chosen model function:

$$e_m(n) = \frac{1}{n^a} + e_\infty$$

- Determine parameters a and e_∞ by nonlinear regression analysis.
- Convert equation to calculate number of samples needed for a desired error rate:

$$n = \sqrt[a]{\frac{1}{e_m(n) - e_\infty}}$$

Experiments

Test method on four data sets:

1. Optical media inspection set from industrial quality inspection, 20 features, 10 classes.
2. Modified NIST set [4] consisting of hand-written digits of size 28x28, 784 features, 10 classes.
3. Artificial Gaussian distributed set $\mathcal{A}$ (Class means table 1, class variances 1, Bayes error probability 39.6%).
4. Set $\mathcal{B}$ like 3., Bayes probability 4.62%.

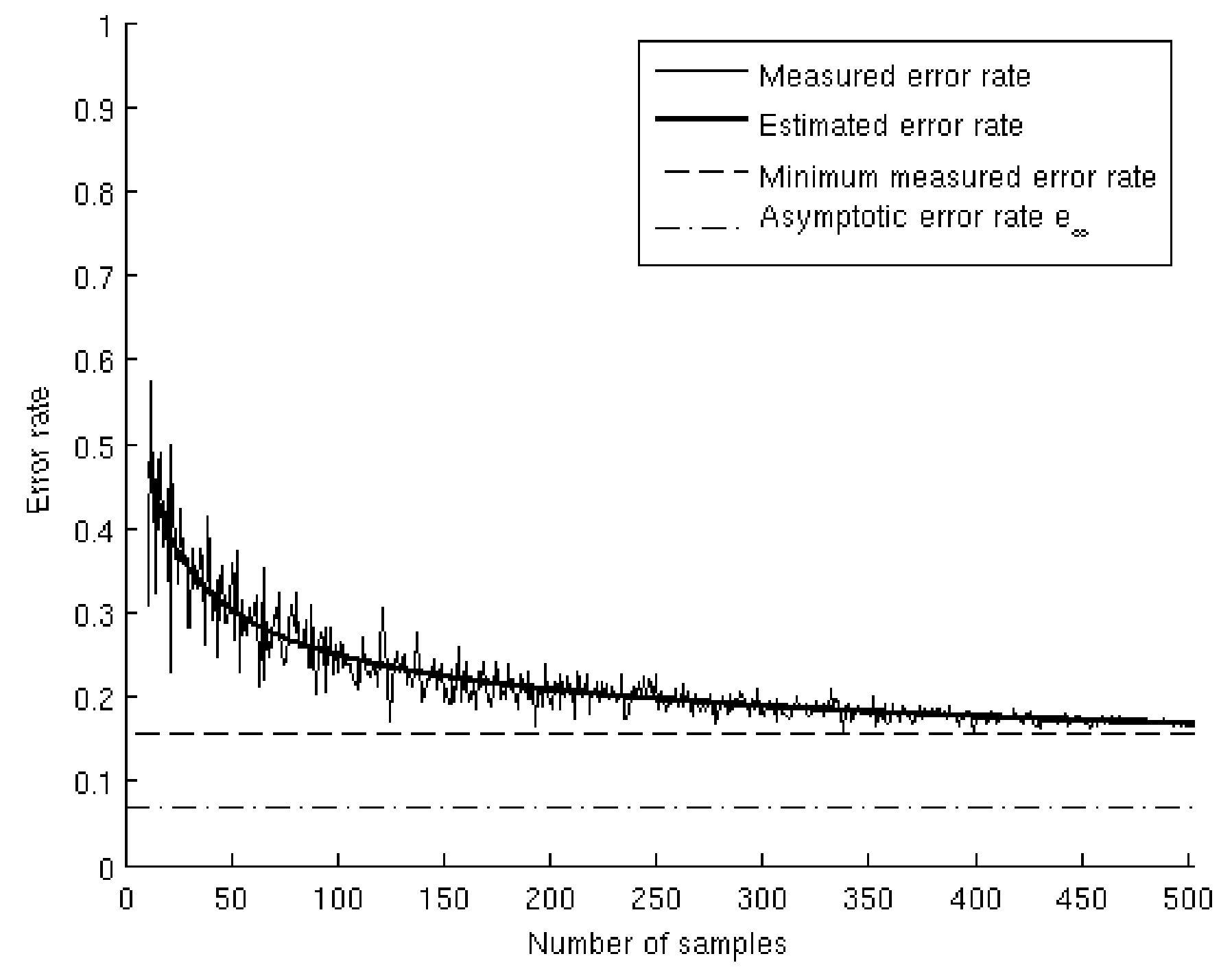
- Use $N/2$ training samples to fit model to measured error rate values
- Compare e_∞ to minimum error rate in real world sets and to Bayes error probability $p_B(e)$ in artificial sets.
- Compare $e(N)$ to extrapolated $e_m(N)$.
- Compare number of samples $\hat{N}$ to reach $e(n)$ to N.

Table 1. Class means for sets $\mathcal{A}$ and $\mathcal{B}$.

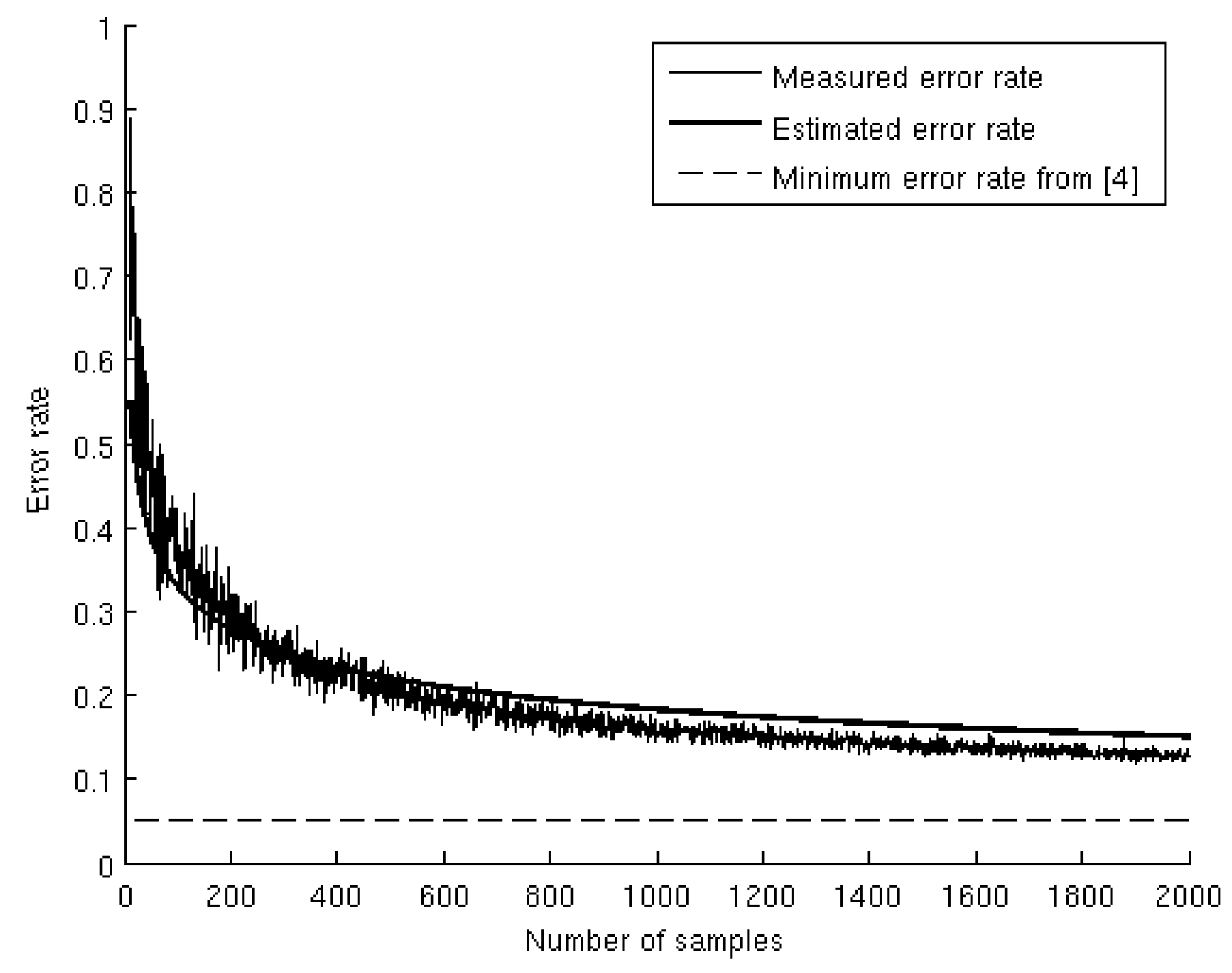
Class	Set $\mathcal{A}$ Feat. 1	Set $\mathcal{A}$ Feat. 2	Set $\mathcal{B}$ Feat. 1	Set $\mathcal{B}$ Feat. 2
ω_1	1	1	2	2
ω_2	1	-1	2	-2
ω_3	-1	1	-2	2
ω_4	-1	-1	-2	-2

Results

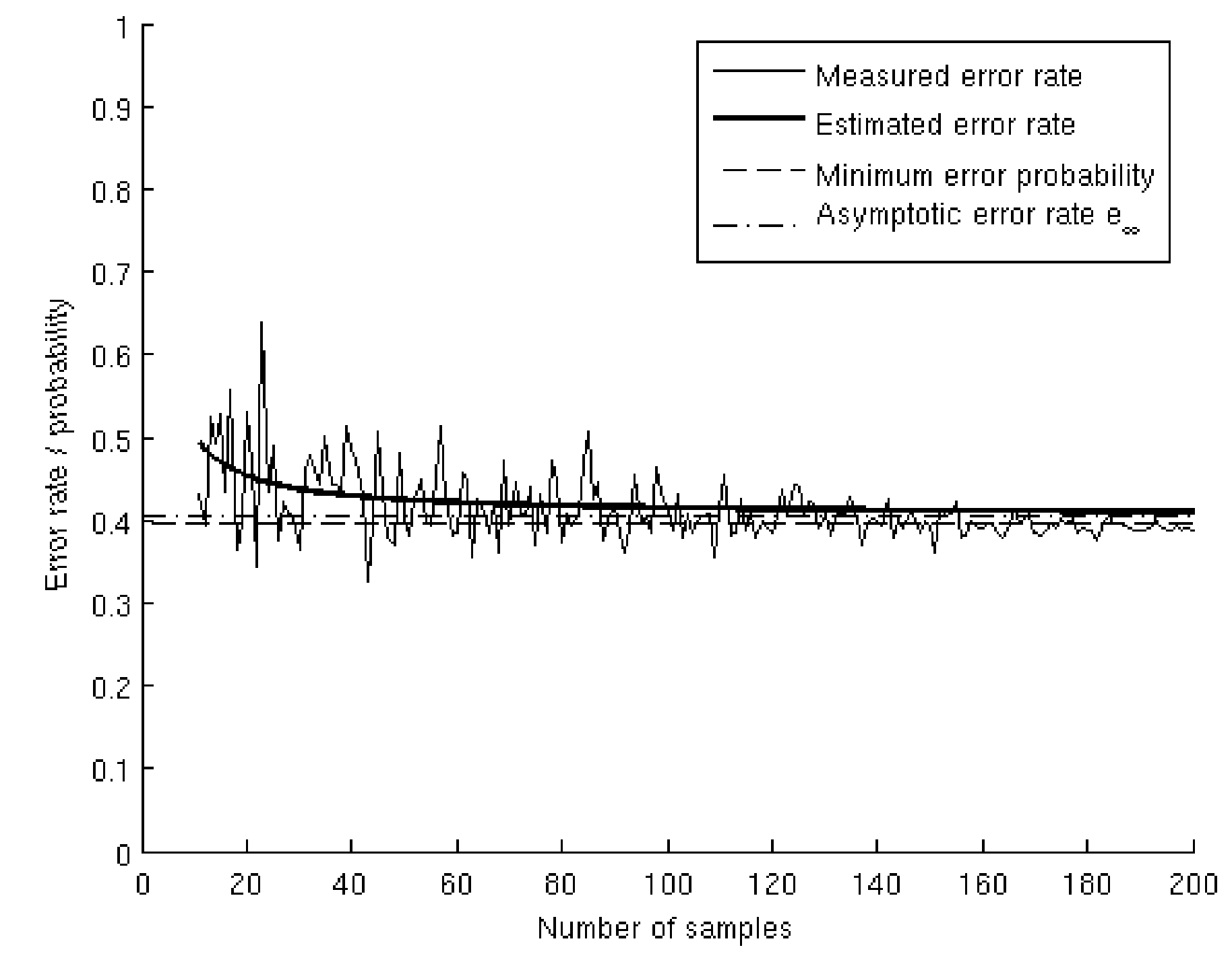
Optical Media Inspection (OMI) data set:



Modified NIST (MNIST) data set:



Artificial Gaussian distributed set $\mathcal{A}$:



Artificial Gaussian distributed set $\mathcal{B}$:

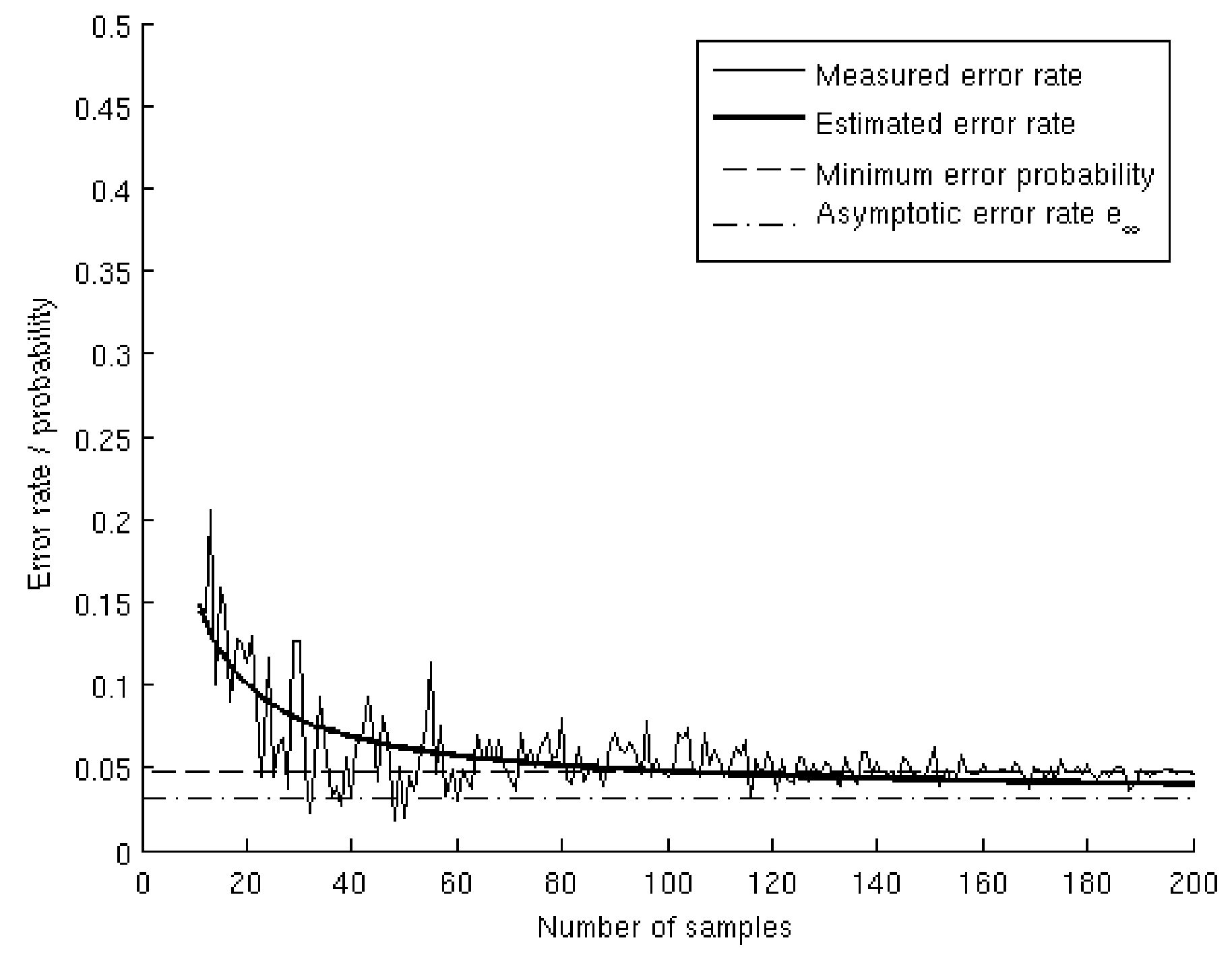


Table 2. Estimation results for all data sets.

Data set	OMI	MNIST	$\mathcal{A}$	$\mathcal{B}$
a	0.3721	0.2021	1.0176	0.8933
N	500	2000	200	200
$\hat{N}$	499	3577	–	43
e_b	15.75%	5.0%	39.6%	4.78%
e_∞	6.91%	-6.47%	40.56%	3.12%
$e(N)$	16.82%	12.64%	38.76%	4.62%
$e_m(N)$	16.79%	15.02%	41.01%	4.0%

e_b = min. error rate for sets 1 and 2 and $e_b = p_B(e)$ for $\mathcal{A}$ and $\mathcal{B}$.

Conclusions

- Method performs best for OMI data set – the targeted kind of problem.
- Parameter e_∞ can be used as quality measure of other values.
- If $\hat{N} < N$ or $e(N) < e_\infty$ samples can be removed from the data set.
- If $\hat{N} = N$ and $e(N) = e_m(N)$ the model can be used to determine error rates beyond N

References

[1] B. V. Dasarathy. *Nearest neighbor classification techniques.* IEEE CS Press, 1990.

[2] R. O. Duda, P. E. Hart, and D. G. Stork. *Pattern Classification.* John Wiley and Sons, 2nd ed., 2001.

[3] H. M. Kalayeh and D. A. Landgrebe. *Predicting the required number of training samples.* IEEE T-PAMI, 5(6):664-667, 1983.

[4] Y. LeCun, L. Bottou, Y. Bengio, and P. Haffner. *Gradient-based learning applied to document recognition.* Procs IEEE, 86(11):2278-2324, 1998.

[5] B. W. Silverman. *Density estimation for statistics and data analysis*, vol. 26 of *Monographs on Statistics and Applied Probability.* Chapman and Hall, London, 1986.