Predicting the benefit of sample size extension in multiclass k-NN classification

Christian Kier
University of Luebeck
Institute for Signal Processing
D-23538 Luebeck, Germany
kier@isip.uni-luebeck.de

Til Aach
RWTH Aachen University
Institute of Imaging and Computer Vision
D-52074 Aachen, Germany

Abstract

In industrial quality inspection obtaining the training data needed for classification problems is still a very costly task. Nevertheless, the classifier quality is crucial for economic success. Thus, the question whether the influence of the training data on the classification error has been fully exploited and enough data has been obtained is very important. This paper introduces a method to answer this question for a specific problem. To be able to make a concrete statement and not only general recommendations, we focus on the k-NN classifier, since it is widely used in industrial implementations. The method is tested on four different multiclass problems: original data from an optical media inspection problem, the MNIST database, and two artificial problems with known probability densities.

1. Introduction

[1] In industrial quality inspection tasks often go beyond only separating good from faulty products. Instead, the occuring defects have to be sorted into different error classes to specifically identify erroneous elements in the production process. This results in multiclass classification problems that have to be solved with an optimal use of resources, i.e. manpower, money and time. A very costly task in the design of a classification system is obtaining the training data with sufficient information for representing the class density distributions. Usually, training data are obtained directly from the production line and not from a separate test environment. Thus, one has to wait until errors occur. No manufacturer would deliberately produce errors in his products only to generate training data for a classification system since it is too expensive and it is not sure that these provoked errors are similar to the errors occuring in the production line.

[1]We gratefully acknowledge the support from Innovationsstiftung Schleswig-Holstein (2004-28H) and from Basler Vision Technologies AG.

Nevertheless, the classification system quality is crucial for economic success. Because of the phenomenon of decreasing error rate with increasing training set cardinality usually as many training samples as possible are obtained. Thereby, the question whether the influence of the training data on the classification error has been fully exploited and enough data has been obtained is very important. In [11] and [12] the authors give formulae for the required number of samples in dependence of the dimensionality to reach an error rate 1.5 times the Bayes error. But usually a designer is more interested in the number of samples beyond which the error rate does not improve significantly. In [9] the data are assumed to be multivariate normally distributed and a criterion estimating the quality of the covariance matrix is developed. Based on this criterion the necessary number of training samples for maximum likelihood classification is predicted. The Gaussian assumption can not always be retained in industrial context. Further attention has been devoted to handling small sample set cardinalities without making statements about set expansion [4, 5, 14].

Usually, an asymptotical error rate $e_\infty = \lim_{n\to\infty} e(n)$ exists for the chosen classifier that can be reached with an infinite number of training samples. But often the error rate $e(n)$ converges fast towards e_∞ and a low number of samples is sufficient for good performance. In place of general recommendations regarding sample size considerations as in [7], [8], or [12], this work tries to give a specific statement, whether the extension of the sample set is worth the additional effort. A specific statement can neither be made for problems in a general manner nor independent of the classifier used. Hence, some assumptions have to be made. Our method analyses a specific classification problem and concentrates on the k-Nearest-Neighbour rule. The k-NN classifier is frequently used throughout industrial classification systems because it is non-parametric, intuitively comprehensible, and provides acceptable classification performance. Downsides of k-NN classification are large computation requirements during operation. They can be overcome by editing and condensing methods [1, 2]. The in-

fluence of editing on the sample size has also been examined [6], but no concrete statement could be made. Hence, a method giving hints about optimal training sample sizes prevents the use of resources to obtain unnecessary training data while assuring optimal classification results in places where a k-NN classifier is used for multiclass classification.

2. Dependency of error rate on sample size

The basic idea of this work is to model the sample size dependent error rate curve $e(n)$ through measurements performed on a specific data set of size N. Starting with N the sample size is decreased by randomly removing samples. For every step the corresponding $e(n)$ is estimated by cross-validation. The obtained values are used to parametrise a model function enabling the classifier designer to extrapolate $e(n)$ beyond N.

An approach to derive the error rate model function $e_m(n)$ is based on estimating the probability of error p_e from the class-conditional probability densities $p(x|\omega_i)$, since the high error rate at small n is based on inaccurate estimation of $p(x|\omega_i)$ by the k-NN classifier [3]. The k-NN probability density estimates are given by [13]:

$$\overline{p(x|\omega_i)} = \frac{k}{n_i \cdot V_k} = \frac{k}{n_i \cdot c_d \cdot r_k^i(x)^d} \tag{1}$$

with n_i being number of samples of class ω_i, c_d the volume of the unit sphere of dimension d, and $r_k^i(x)$ the distance to the k^{th} nearest neighbour of class ω_i. Thus, V_k is the volume of the hypershere around x spanning over k nearest neighbours. Let c be the number of classes, $\mathcal{R}_i$ the decision region for ω_i, $P(\omega_i)$ the prior and $P(\omega_i|x)$ the posterior probabilities. The overall error probability p_e is given by (mutatis mutandis to [3]):

$$\begin{aligned} p_e &= \sum_{j=1}^{c} \int_{\mathcal{R}_j} \sum_{i=1, i\neq j}^{c} P(\omega_i|x)\, dx \\ &= \sum_{j=1}^{c} \int_{\mathcal{R}_j} \left[1 - P(\omega_j|x)\right]\, dx \\ &\overset{Bayes}{=} \sum_{j=1}^{c} \int_{\mathcal{R}_j} \left[1 - \frac{p(x|\omega_j)P(\omega_j)}{\sum_i p(x|\omega_i)P(\omega_i)}\right] dx \end{aligned}$$

To estimate p_e, (1) is used to estimate the prior probabilities. Even with the assumption of equal prior probabilities and equal class cardinalities this expression only simplifies to

$$\overline{p_e(n)} = \sum_{j=1}^{c} \int_{\mathcal{R}_j} \left[1 - \frac{\frac{1}{n_j r_k^j(x)^d} P(\omega_j)}{\sum_i \frac{1}{n_i r_k^i(x)^d} P(\omega_i)}\right] dx$$

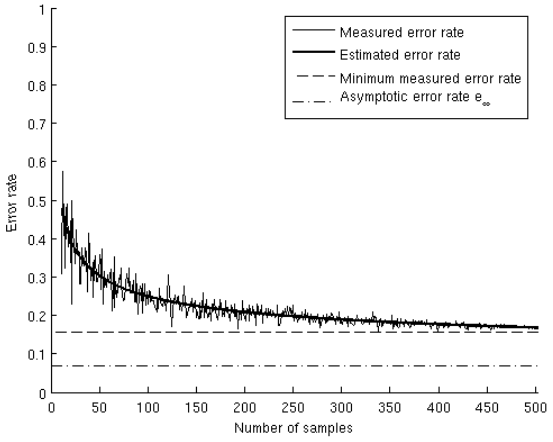

Figure 1. Measured error rate and fitted error curve for optical media inspection data set.

$$\overset{\substack{P(\omega_i)=P(\omega_j)\\ n_i=n_j}}{=} \sum_{j=1}^{c} \int_{\mathcal{R}_j} \left[1 - \frac{n}{n_j} \frac{\frac{1}{r_k^j(x)^d}}{\sum_i \frac{1}{r_k^i(x)^d}}\right] dx. \tag{2}$$

Due to the occurrence of $r_k^j(x)$ this expression is hard to handle. Thus, an artificial model function is used to estimate the error rate. Since with growing n the coefficient $\frac{n}{n_i}$ should remain almost constant and $r_k^j(x)$ becomes smaller the simplest model matching the character of equation 2 and simultaneously yielding the best results would be:

$$e_m(n) = \frac{1}{n^a} + e_\infty. \tag{3}$$

The values obtained through the measurements with decreasing n are used to determine the two parameters a and e_∞ with nonlinear regression analysis. A good starting value for a is $\sqrt{2}$. For e_∞ the error rate at the maximum available n should be chosen as starting value.

When all parameters have been determined, the estimated asymptotic error rate e_∞ tells the classifier designer how good the classifier could asymptotically become. One can also calculate $e_m(n)$ at values $n > N$ to get an indication whether obtaining additional samples is worth the effort. The expression in (3) could also be used to calculate the number of samples needed for a desired error rate:

$$n = \sqrt[a]{\frac{1}{e_m(n) - e_\infty}}. \tag{4}$$

3. Experiments

The method has been tested on four different data sets – two real world problems and two artificially generated prob-

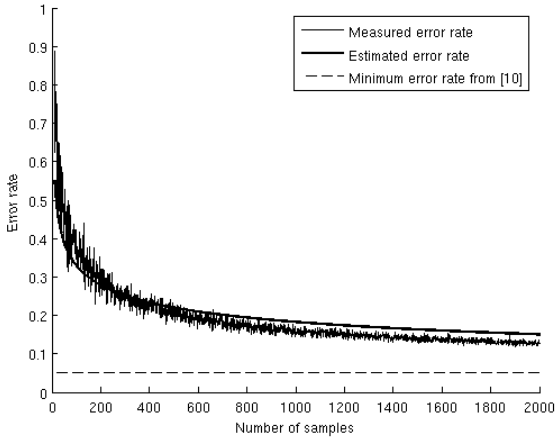

Figure 2. Measured error rate and fitted error curve for the MNIST data set.

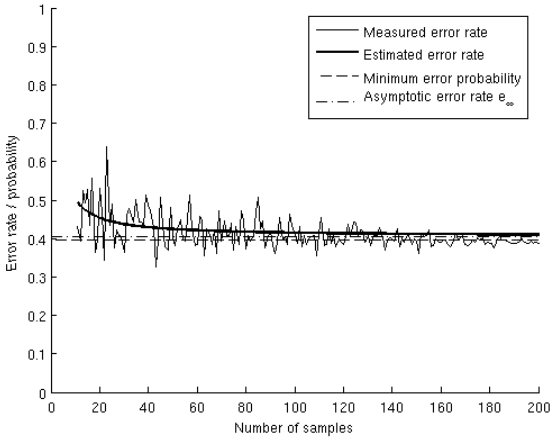

Figure 3. Measured error rate and fitted error curve for the artificial data set $\mathcal{A}$.

lems. For each data set, the samples have been split randomly in equally sized train and test subsets. To be able to assess the method's quality $N/2$ training samples have been used to fit (3) to the measured error rate values $e(n)$. The asymptotic error rate $e_\infty$is compared to the minimum error rate $\min_{n\leq N} e(n)$ in sets 1 and 2 and with the Bayes error probability $p_B(e)$ in sets 3 and 4. Furthermore, the measured error rate $e(N)$ is compared to the extrapolated error rate $e_m(N)$ and the number of samples $\hat{N}$ to reach $e(N)$ is compared to N. The results for all data sets are shown in table 2. All error rate curves $e(n)$ have also been obtained for different values of k but the results differed only marginally. Thus, only the results for $k = 3$ are shown.

The following data sets have been used:

1. Optical Media Inspection (OMI) set
   This data set comes from industrial quality inspection in optical media production. From the obtained images 20 features are calculated. The set consists of 1000 samples in 10 classes, i.e. $N = 500$ (see fig. 1).

2. Modified NIST (MNIST) set
   This data set consists of hand-written digits from 0 to 9 originating from zip code images [10]. The images are of size 28x28 pixels resulting in 784 features. In [10] the minimum 3-NN error rate is given as 5.0%. The original data set consists of 60000 train images and 10000 test images but here only 4000 samples are used, i.e. $N = 2000$ (see fig. 2).

3. Artificial Gaussian distributed set $\mathcal{A}$
   This data set is artificially generated to be able to calculate the Bayes error probability for comparison. It is two-dimensional and consists of 400 samples in four classes, i.e. $N = 200$. The class distributions are Gaussian and differ only in their means (see table 1). They all have variance $\sigma^2 = 1$. The Bayes error probability of this set is $p_B(e) = 39.6\%$ (see fig. 3).

Table 1. Feature-wise class means for data sets $\mathcal{A}$ and $\mathcal{B}$.

	Set $\mathcal{A}$		Set $\mathcal{B}$	
Class	Feat. 1	Feat. 2	Feat. 1	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

4. Artificial Gaussian distributed set $\mathcal{B}$
   This data set is similar to set $\mathcal{A}$ except for the class means (see table 1) resulting in a Bayes error probability $p_B(e) = 4.78\%$ (see fig. 4).

For sets 3 and 4 values of $e(n)$ below the corresponding values of $p_B(e)$ occur due to the fact that subsets are chosen randomly resulting in possibly "better" class distributions.

4. Discussion and Conclusions

For the OMI data set the method works pretty well. The error rate at 500 samples is almost exactly predicted from the first 250 samples. The number of samples needed to reach $e(N) = 16.82\%$ is exactly predicted as well. Noticeable is the rather strong deviation of $e_\infty$from the minimum measured error rate. But since this is a real world problem

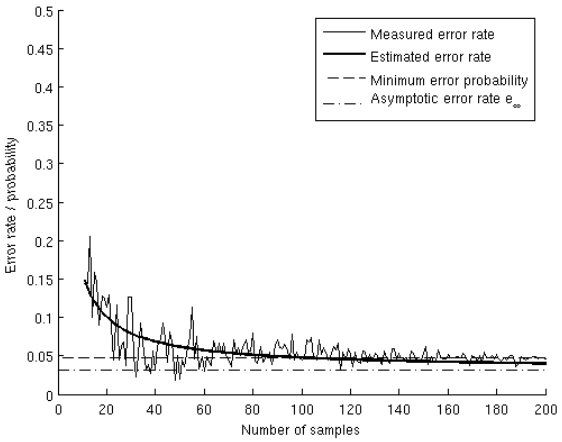

Figure 4. Measured error rate and fitted error curve for the artificial data set $\mathcal{B}$.

where no more samples are available no statement can be made concerning the accuracy of e_∞.

With the MNIST data set an inherent shortcoming of the method becomes evident. The predicted value of e_∞ is negative. Nevertheless, the measured error rate curve is well approximated with $e_m(N)$ being slightly larger than $e(N)$. The number of needed samples $\hat{N}$ is misleading though, since $e_m(N)$ can already be reached with 1200 samples.

For the artificial set $\mathcal{A}$ it is not possible to calculate $\hat{N}$ since $e(N) < e_\infty$. For set $\mathcal{B}$ N and $\hat{N}$ differ quite strongly. But unlike the MNIST data set this shows a strength of the method since $\hat{N} < N$. The figures 3 and 4 show that the minimum error rate has already been reached at lower values of n. This tells the designer that is even possible to remove samples from the training set. The values of e_∞ are moderately accurate: Slightly too high for set $\mathcal{A}$ and too low for set $\mathcal{B}$. The approximation of the error rate with N samples is pretty good for set $\mathcal{A}$. For set $\mathcal{B}$ the value is too low. This is due to the fact, that the minimum error rate has already been reached, but $e_m(n)$ is still further decreasing.

It is noticeable that the presented method is performing best for the OMI data set – a problem for which it is intended to be used. This result can still be improved if the total number of available samples is used for parameter determination. From the two calculated parameters e_∞ can only be used as some kind of quality measure for the other values. If it is negative or differs strongly from e_b, the other values are to be treated with care. The parameter a gives a hint how fast the curve is decreasing.

To summarise, if $\hat{N} < N$ or $e(N) < e_\infty$ samples can be removed from the training set. If $\hat{N} \approx N$ and $e(N) \approx e_m(N)$ the error rate model can be used to determine error rates beyond N.

Table 2. Estimation results for all four 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 $\mathcal{B}$.

References

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

[2] P. A. Devijver and J. Kittler. *Pattern recognition: A statistical approach*. Prentice-Hall, London, 1982.

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

[4] R. P. Duin. Classifiers in almost empty spaces. In *Procs ICPR*, volume 2, pages 2001–2007, Barcelona, 2000.

[5] R. P. W. Duin. Small sample size generalization. In *Procs* 9^{th} *Scandinavian Conference on Image Analysis*, Uppsala, Sweden, June 1995.

[6] F. J. Ferri and E. Vidal. Small sample size effects in the use of editing techniques. In *Procs ICPR*, pages 607–610, The Hague, The Netherlands, 1992.

[7] K. Fukunaga and R. R. Hayes. Effects of sample size in classifier design. *IEEE T–PAMI*, 11(8):873–885, August 1989.

[8] A. K. Jain and B. Chandrasekaran. *Dimensionality and Sample Size Considerations in Pattern Recognition Practice*, volume 2 of *Handbook of Statistics*, chapter 39, pages 835–855. North-Holland, 1982.

[9] H. M. Kalayeh and D. A. Landgrebe. Predicting the required number of training samples (for remotely sensed image data based on covariance matrix estimate quality criterion of normal distribution). *IEEE T–PAMI*, 5(6):664–667, 1983.

[10] Y. LeCun, L. Bottou, Y. Bengio, and P. Haffner. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324, November 1998.

[11] S. Raudys and V. Pikelis. On dimensionality, sample size, classification error, and complexity of classification algorithms in pattern recognition. *IEEE T–PAMI*, 2(3):242–252, 1980.

[12] S. J. Raudys and A. K. Jain. Small sample size effects in statistical pattern recognition: Recommendations for practitioners. *IEEE T–PAMI*, 13(3):252–264, March 1991.

[13] B. W. Silverman. *Density Estimation for Statistics and Data Analysis*, volume 26 of *Monographs on Statistics and Applied Probability*. Chapman and Hall, London, 1986.

[14] M. Skurichina and R. P. W. Duin. Stabilizing classifiers for very small sample sizes. In *Procs ICPR*, volume 2, pages 891–896, 1996.