The Evaluation of Feature Extraction Criteria Applied to Neural Network Classifiers

W. Utschick P. Nachbar C. Knobloch A. Schuler J.A. Nossek

Institute of Network Theory and Circuit Design
Technical University of Munich
Arcisstr. 21, 80333 Munich, Germany

Abstract

Feature extraction is a crucial part of classification procedures. In this paper we present an approach, how to utilize feature extraction criteria to predict the potential efficiency of a neural network classifier. Statistical and geometrical criteria are introduced for analysis. The complete system of our research consists of a class of generalized Hough–Transformations for feature extraction and a subsequent neural network. The neural network performs the classification based on respective features. For an example we concentrated on a pattern recognition problem — the classification of handwritten numerals. As a result of our work we assign two feature extraction criteria to the employed network for a significant estimation of its efficiency.

Keywords: *feature extraction, classification, statistical criteria, Hough–Transformation, neural networks, prediction.*

1 Introduction

A wide variety of approaches has been taken towards the task of classification. Given a pattern space $\mathcal{X}$ consisting of M mutually exclusive sets $\mathcal{X} = \mathcal{C}_1 \cup \mathcal{C}_2 \cup \cdots \cup \mathcal{C}_M$ with each of $\mathcal{C}_j, \forall_j \in \Lambda = \{1, 2, \cdots, M\}$ representing a set of specified patterns called a class, e.g. $M = 10$ for the problem of numeral character recognition, we may simple regard a classifier as a unit that receives an input sample and outputs a label j.

Two main directions of research in pattern recognition are conventional approaches (e.g. statistical classifiers) and neural networks. Whereas conventional classifiers may be set up of at least two subsystems, a feature extractor and the actual classifier, neural networks inherently combine statistical techniques in a more implicit way. Because of the "unconscious" kind of learning in neural networks they apparently need no comparable efforts in feature extraction as conventional approaches do. However, applying neural networks to the raw input data without any feature extraction or using prior knowledge requires a huge amount of training samples and increases the problem of learning and generalization. For the special tasks of pattern recognition the combination of neural approaches with feature extraction strategies and the use of prior knowledge has become an alternative to existing systems [1, 2]. In this article we introduce a neural network classifier combined with an extended Hough–Transformation for feature extraction applied to the problem of single character recognition.

In the subsequel of this paper, we briefly introduce three known criteria for comparison of feature extraction methods which are inherently related to particular classification procedures [3, 4]. Before applying these criteria to real data in the example of handwritten characters and analyzing their predictive power for neural networks, we derive how to measure the correlation between feature extraction and classifier performance when classifying multiple classes of patterns. The objective of this work is to demonstrate the significance of refering feature extraction criteria to neural classification procedures.

2 Criteria for feature extraction

An obvious criterion for feature extraction is the probability of classification errors. Unfortunately, the error probability of a classifier is generally very difficult to calculate. Although the observed error ratio of a classification procedure is a sufficient estimate of the error probability, its calculation requires high computational costs. This is especially not feasible for any optimization algorithm which iteratively searches for an optimal feature extraction (selection) algorithm. Therefore, in the following subsections we consider two alternative types of criteria for evaluating feature extraction, each of them related to respective classification procedures.

2.1 Statistical criteria

Consider two classes of patterns $\mathcal{C}_1$ and $\mathcal{C}_2$, with probability density functions $p_1(\mathbf{x})$ and $p_2(\mathbf{x})$ for $x \in \mathcal{X} = \mathcal{C}_1 \cup \mathcal{C}_2$. A measure for discrimination of the two classes is the so called divergence Q_d [3, 4]:

$$Q_d = \int_{\mathcal{X}} p_1(\mathbf{x}) \log \frac{p_1(\mathbf{x})}{p_2(\mathbf{x})} d\mathbf{x} + \int_{\mathcal{X}} p_2(\mathbf{x}) \log \frac{p_2(\mathbf{x})}{p_1(\mathbf{x})} d\mathbf{x} \,. \tag{1}$$

The divergence is based on the log-likelihood ratio and may be interpreted as a measure of information for discrimination of the two classes. By $\frac{1}{8} \exp(-Q_d)$ the divergence provides a lower bound of the Bayes error probability.

A second statistical measure is the Bhattacharyya coefficient or the sometimes more convenient Bhattacharyya distance Q_B [3, 4]:

0-8186-7128-9/95 $4.00 © 1995 IEEE

$$Q_B = -\log \int_{\mathcal{X}} (p_1(\mathbf{x}) p_2(\mathbf{x}))^{\frac{1}{2}} d\mathbf{x} \,. \qquad (2)$$

An important property of the Bhattacharyya distance is that $\frac{1}{2}\exp(-Q_b)$ is an upper bound on the error probability of a corresponding Bayes classifier.
For the assumption of Gaussian densities, with $\bar{\mathrm{x}}_{1,2}$ and $\mathrm{C}_{1,2}$ representing the expectation values and the correlation matrices of the classes $\mathcal{C}_1$ and $\mathcal{C}_2$, the introduced criteria are given by:

$$\begin{aligned} Q_d &= \frac{1}{2}(\bar{\mathrm{x}}_1 - \bar{\mathrm{x}}_2)^T (\mathrm{C}_1^{-1} + \mathrm{C}_2^{-1})(\bar{\mathrm{x}}_1 - \bar{\mathrm{x}}_2) \\ &\quad + \frac{1}{2}\mathrm{trace}(\mathrm{C}_1^{-1} - \mathrm{C}_2^{-2})(\mathrm{C}_2^{-1} - \mathrm{C}_1^{-1}) \qquad (3) \\ Q_B &= \frac{1}{8}(\bar{\mathrm{x}}_1 - \bar{\mathrm{x}}_2)^T \left(\frac{\mathrm{C}_1 + \mathrm{C}_2}{2}\right)^{-1} (\bar{\mathrm{x}}_1 - \bar{\mathrm{x}}_2) \\ &\quad + \frac{1}{2}\log(\left|\frac{\mathrm{C}_1 + \mathrm{C}_2}{2}\right| |\mathrm{C}_1|^{-\frac{1}{2}} |\mathrm{C}_2|^{-\frac{1}{2}}) \qquad (4) \end{aligned}$$

2.2 Geometrical criteria

Applying adequate transformations to feature extraction and classification is based on an intuitive notion [3, 4]: It is desirable to have a transformation such that the separation of different classes is improved after transformation. A suitable criterion for separability is the geometrically induced ratio of interclass–to–intraclass distances given by

$$Q_{g'} = \frac{\mathrm{trace}\ \mathbf{S}_b}{\mathrm{trace}\ \mathbf{S}_w} \,. \qquad (5)$$

The so called "within–class" and "between–class" scatter matrices are defined as follows:

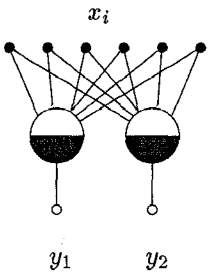
$$\begin{aligned} \mathbf{S}_w &= \sum_{\Lambda} p_j \left(\frac{1}{\sharp|\mathcal{C}_j|} \sum_{\mathcal{C}_j} (\mathrm{x}_i - \bar{\mathrm{x}}_j)(\mathrm{x}_i - \bar{\mathrm{x}}_j)^T \right) ,(6) \\ \mathbf{S}_b &= \frac{1}{2} \sum_{\Lambda} \sum_{\Lambda} p_j p_k (\bar{\mathrm{x}}_j - \bar{\mathrm{x}}_k)(\bar{\mathrm{x}}_j - \bar{\mathrm{x}}_k)^T \,. \qquad (7) \end{aligned}$$

For optimal performance the scatter matrices are transformed by means of an adequate normalization procedure. The modified version of the geometrical criteria is called Q_g. For optimal performance the scatter matrices are equally transformed. According to its definition the criteria is totally compatible to multiple class problems.

3 Criteria for multiple class discrimination

When analyzing the correlation of feature extraction criteria with the expected error probability of a respective classification approach two questions arise. First how to deal with multiple class problems, if we intend to utilize criteria, which are generally related only to two class problems like the divergence or the Bhattacharyya distance, and second how to measure a correlation between criteria and real classifiers. It is to remark, that there exist extensions of the statistical criteria for multiple class in the literature [3, 4]. In the following a more pragmatical and appropriate strategy is outlined.

$$\Theta = \{\mathcal{C}_1, \mathcal{C}_2, \mathcal{C}_3, \mathcal{C}_4\}$$



	y_1	y_2
$\mathcal{C}_1$	-1	-1
$\mathcal{C}_2$	-1	$+1$
$\mathcal{C}_3$	$+1$	-1
$\mathcal{C}_4$	$+1$	$+1$

⇩

$$y_2 \Rightarrow \Theta = \{\mathcal{C}_1, \mathcal{C}_3\} \cup \{\mathcal{C}_2, \mathcal{C}_4\}$$

Figure 1: Example of a neural network: the binary output y_2 of the architecture corresponds to a particular dichotomy of Θ

For the problem of multiple classes we have to consider that neural classifiers are in principle related to two class problems. If classifiers produce a binary valued output the relation to two class problems is obvious. The binary output code directly implies a dichotomy of $\mathcal{X} = \mathcal{C} \cup \mathcal{X} \backslash \mathcal{C}$ in $\mathcal{C}$ and its complement according to each output line of the system. For example, considering the discrimination of four classes $\Theta = \{\mathcal{C}_1, \mathcal{C}_2, \mathcal{C}_3, \mathcal{C}_4\}$ of patterns by a neural network with totally two output lines in Figure 1. The output y_2 of the classifier obviously represents a particular dichotomy of the four classes, e.g. $\mathcal{C} = \{\mathcal{C}_2, \mathcal{C}_4\}$ corresponding with its complement $\bar{\mathcal{C}} = \mathcal{X} \setminus \mathcal{C} = \{\mathcal{C}_1, \mathcal{C}_3\}$.
For applying the given criteria for feature extraction to multiple class discrimination we introduce an extension by calculating the mean of the respective criteria applied to a significant subset $\mathcal{S} \subseteq 2^{\Theta}$ of dichotomies (two class problems) of the corresponding classes of patterns. With the superset 2^{Θ} consisting of all possible dichotomies of Θ the modified criteria for feature extraction are given by

$$Q_{\cdot} = \frac{1}{\sharp|\mathcal{S}|} \sum_{\mathcal{S}} Q_{\cdot,i} \,. \qquad (8)$$

4 Relation to empirical criteria

For relating derived criteria to predict the expected efficiency of the corresponding classifier a regression approach has been applied. Therefore, the calculation of the given criteria, as well as the collection of empirical data of the neural network, has to be performed according to a sufficient large number of different feature extraction settings. Subsequently, from a least squares point of view, regression techniques supply the best approximation or prediction of the classifier's empirical quality based on the feature extraction criteria. The correlation coefficient $\rho_{\cdot}$ is given by

$$\rho_{\cdot} = \frac{\sum(Q_{\cdot} - \bar{Q}_{\cdot})(Q_e - \bar{Q}_e)}{\left(\sum(Q_{\cdot} - \bar{Q}_{\cdot})^2 \sum(Q_e - \bar{Q}_e)^2\right)^{\frac{1}{2}}}, \qquad (9)$$

with the empirical criterion Q_e, derived from the mean of the observed error ratio ϵ_{err} of the classifier after training and testing the neural network. Due to the random initialization of the learning algorithm each measure of Q_e is averaged by r training and testing runs:

$$Q_e = \left(\frac{1}{\sharp|\mathcal{S}| \cdot r} \sum_{\mathcal{S}} \epsilon_{err}\right)^{-1} \qquad (10)$$

5 Example

The subject of our experiment has been a neural network classifier for recognizing handwritten numerals. The data we use are provided by the CEDAR CDROM Database [5]

Before supplying to the classifier the raw data is transformed by an extended version of the Hough–Transformation.

5.1 Hough–Transformation

For the purpose of feature extraction we implemented an modified version of the Hough–Transformation. Instead of matching the bitmaps of input characters with a significant bundle of lines and circles, a variety of stripes or/and ellipsoids are applied to the bitmap. By segmentation and rotation of parallel stripes an appropriate set of "linear"features is spread over a pattern.

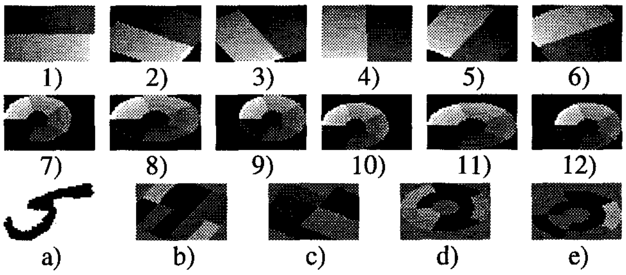


Figure 2: A mixture of "linear" and "elliptical" prototypes (1–12) and the pattern "5" after Hough–Transformation (a–e)

An additional type of features is derived by matching with a mixture of concentric ellipsoids and its partitioning in particular sectors as shown in Figures 2.1–2.12. Due to the constant number of inputs of neural networks and compatibility of results the number of features is fixed, i.e. totally 40 particular features are extracted by a combination of concatenated stripes or sectors. The

- resolution of the grid
- the width of the stripes and their orientation
- the centers of the ellipsoids
- their number of sectors and radii,

as well as the possibility of a mixture enables a large variety of different appearances of the transformation. Figure 2a–2e depicts the decomposition of the character "5" by the Hough–Transformation consisting of a mixture of "linear" and "elliptical" prototypes given in 2.1–2.12.

In this paper we used a set of 30 different Hough–Transformations for the extraction of significant features of handwritten numerals. The modified Hough–Transformation serves as a flexible tool covering pure pattern matching techniques as well as structural techniques. Therefore, the presented Hough–Transformation may be representative for a variety of feature extraction approaches.

5.2 Neural network classifier

After feature extraction we employed a neural network for classifying the handwritten characters. Figure 3 shows the committee–machine like architecture of the applied neural network [6, 7, 8]. The network consists of a single layer perceptron (SLP) combined with a fixed Boolean function (B) in the second layer. The multiple binary output of the network is inherently related to a coding scheme of the different classes of pattern, e.g. 1 out of 10. The training of the classifier was carried out by an extension of the Madaline I algorithm [7, 8], the so called MadaTron [9, 10].

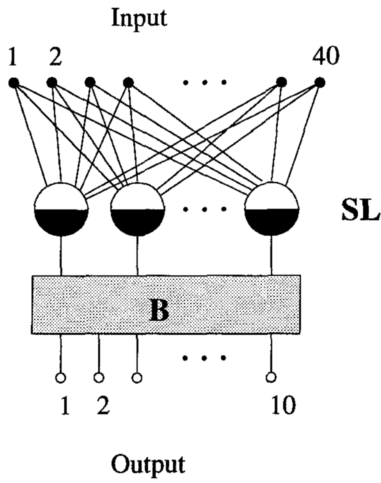
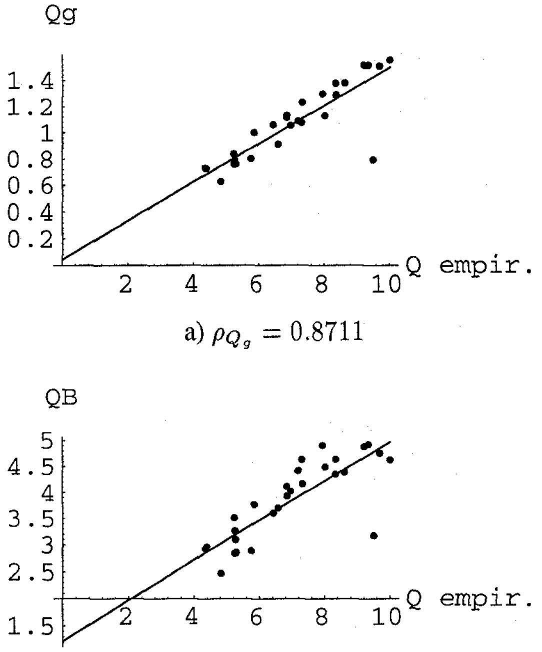


Figure 3: A two–layer neural network architecture consisting of a SLP and a Boolean function

6 Results

The results in this paper are based on 1000 randomly drawn handwritten numerals provided by the CEDAR CDROM Database. By selecting different prototypes of features we created 30 different versions of the Hough–Transformation. According to the neural network classifier the approach is focused on a subset $\mathcal{S}$ of $\sharp|\mathcal{S}| = 77$ out of $\frac{2^{10}}{2}$ significant dichotomies of classes of patterns, inclusive the elementary 1 out of 10 codes. With regard to the random initialization of the MadaTron algorithm every training and testing phase of the network was repeated $r = 3$ times. For each $\rho_{\cdot}$ the respective criterion $Q_{\cdot}$ and the empirical



a) $\rho_{Q_g} = 0.8711$

b) $\rho_{Q_B} = 0.8214$

Figure 4: Correlation of empirical criterion Q_e with a) interclass–intraclass distances Q_g and b) Bhattacharyya criterion Q_B

criterion Q_e is calculated repeatedly, i.e. for all 30 feature extracting procedures (different modifications of the Hough–Transformation). For each assignment of the Q_e, totally 3×77 sequences of training and testing are performed.

Additionally, we assumed Gaussian distribution of data for applying statistical criteria.

Figure 4 shows the results of our experiments: the correlations of the introduced feature extraction criteria with the efficiency of the applied neural network classifier based on a set of different feature extraction methods. The highest rate of regression ($\rho_{Q_g} = 0.8711$) is achieved by the geometrically induced interclass–to–intraclass distance Q_g (Figure 4a). A comparison to the statistical criteria shows that the Bhattacharyya distance Q_B equally produces an acceptable result (Figure 4b) whereas the divergence reveales no significant correlation. Hence, the interclass–to–intraclass distance Q_g and the Bhattacharyya distance is proven to be the most adequate criteria for the prediction of the classifier performance based on the applied type of feature extraction approaches.

7 Conclusion

In this contribution we demonstrate how to relate feature extraction criteria to a neural network classifier. Statistical criteria are directly denoted to statistical classification approaches (e.g. Bayes Classifier). For neural networks there exist no a priori criteria for prediction of efficiency. The application of neural networks to discrimination may be interpreted as an approximation of the optimal Bayes classification rule. However, it is not clear how to measure the affinity to appropriate feature extraction. Which criteria are highly correlated to separability and internal representation of data within a neural network approach?

We introduced three different criteria for feature extraction. For dealing with multiple class problems our approach is based on a subset of two–class problems. Two–class problems — dichotomies of respective classes — are implicitely related to the binary output property of neural networks.

For our investigation we used a classification procedure consisting of two stages. After feature extraction (Hough–Transformation) a committee–machine like neural network realizes the real classifier unit. For the regression approach the calculation of criteria has to be performed for a variety of different feature extraction versions. Thereby, the empirical criterion corresponds to the observed error ratio of the neural network after training.

The result of our experiment is based on the CEDAR CDROM Database consisting of single characters of handwritten numerals. It is shown that there is a close relation between the factual efficiency rates of the neural classifier and the values of two particular feature extraction criteria. In a parallel work we use the identified criteria for the optimization of a single character classifier.

References

[1] Y. LeCun. Backpropagation Applied to Handwritten Zip Code. *Neural Computation*, 1:541–551, 1989.

[2] Y. LeCun. *Generalization and Network Design Strategies*. North-Holland, Amsterdam, 1989.

[3] P.A. Devijver and J. Kittler. *Pattern Recognition: A Statistical Approach*. Prentice Hall, 1982.

[4] K. Fukunaga. *Introduction to Statistical Pattern Recognition*. Academic Press, 1972.

[5] Center of Excellence for Document Analysis and Recognition (CEDAR) Backprop applications — 6, Buffalo, NY 14260. *Database of Handwritten Cities, States, ZIP Codes, Digits and Alpha Characters*, CEDAR CDROM edition, 1992.

[6] N. Nilsson. *The Mathematical Foundation of Learning Machines*. Morgan Kaufmann Publishers San Mateo, California, 1990.

[7] B. Widrow. ADALINE and MADALINE – 1963. In *First International Conference on Neural Networks*, volume 1, pages 143–158. IEEE, 1987.

[8] B. Widrow and M.A. Lehr. 30 Years of Adaptive Neural Networks: Perceptron, Madaline, and Backpropagation. *IEEE Proc.*, 78(9):1415–1441, 1990.

[9] P. Nachbar, J. Strobl, and J.A. Nossek. The generalized adatron algorithm. In *International Symposium on Circuits and Systems*, volume 4, pages 2152–2156. IEEE, 1993.

[10] P. Nachbar. *Entwurf robuster neuronaler Netze*. PhD thesis, Technical University Munich, Institute for Network Theory and Circuit Design, 1994.