Trainable fusion rules. II. Small sample-size effects

Šarūnas Raudys*

Institute of Mathematics and Informatics, Akademijos 4, Vilnius 08663, Lithuania

Received 30 January 2003; accepted 2 January 2006

Abstract

Profound theoretical analysis is performed of small-sample properties of trainable fusion rules to determine in which situations neural network ensembles can improve or degrade classification results. We consider small sample effects, specific only to multiple classifiers system design in the two-category case of two important fusion rules: (1) linear weighted average (weighted voting), realized either by the standard Fisher classifier or by the single-layer perceptron, and (2) the non-linear Behavior-Knowledge-Space method. The small sample effects include: (i) training bias, i.e. learning sample size influence on generalization error of the base experts or of the fusion rule, (ii) optimistic biased outputs of the experts (self-boasting effect) and (iii) sample size impact on determining optimal complexity of the fusion rule. Correction terms developed to reduce the self-boasting effect are studied. It is shown that small learning sets increase classification error of the expert classifiers and damage correlation structure between their outputs. If the sizes of learning sets used to develop the expert classifiers are too small, non-trainable fusion rules can outperform more sophisticated trainable ones. A practical technique to fight sample size problems is a noise injection technique. The noise injection reduces the fusion rule’s complexity and diminishes the expert’s boasting bias.

Keywords: Generalization error; Complexity; Learning set; Neural network ensembles; Multiple classifiers systems; Classifier combination; Fusion

1. Introduction. Definitions

To develop a pattern recognition system one needs to use a design set (DS) and a test set (TS). The DS is used to determine system type, an architecture and parameters of a decision-making algorithm. The TS should be used only once, for final evaluation of the system’s performance. If TS is used to select the best classifier and feature subset, to find the optimal number of hidden units of multilayer perceptron (MLP) classifier and optimal number of training epochs, then TS becomes a part of the training set/design set. TS based estimates of the system’s performance become biased. The bias could become substantial in small sample situations. If hold-out error counting estimates (cross validation) are used to evaluate performances of several versions of the algorithm, apparent classification error after model selection decreases by $\sqrt{\frac{P_{\text{test}}(1 - P_{\text{test}})}{n_{\text{test}}}}$ or even more (Raudys, 2001, chap. 6). Above, $P_{\text{test}}$ denotes the average classification error and $n_{\text{test}}$ is the number of vectors in TS.

If one wishes to have unbiased estimates of classification error rate, one is obliged to split DS into learning set (LS) and validation set (VS). Learning (training) set is used to train the classifier. The VS is used to evaluate training bias and estimate real performance of different versions, sets of features, etc. If the validation set is used to select some parameters of the pattern recognition algorithm, performance estimates evaluated on VS also become optimistically biased. Therefore, to train fusion rule in multiple classifiers system (MCS) design one needs a second independent learning set. Thus, strictly speaking, the design set should be partitioned into three parts: (a) learning set No 1, (b) validation set and (c) learning set No 2. No one knows how to partition the DS optimally. Very often partitioning is not done. In spite of partitioning, habitually an original undivided design set is utilized to train the classifier, to select its parameters and to train the fusion rule. It is erroneous practice. Sometimes, certain mathematical formulas are used to overcome the training bias. We need to evaluate the model selection bias too. So, this strategy is not completely optimal too.

A vast amount of theoretical analysis is performed to evaluate the training bias (classifiers’ adaptation to training set), which can be found in reviews by Raudys and Young.
In finite learning set size cases, while training the linear single layer perceptron (SLP) or adaptive linear element ADALINE one can also obtain the standard Fisher DF (Haykin, 1999). The Fisher DF is one of the best classification rules ever suggested. If \( N_2 = N_1 = N \), the expected value of conditional probability of misclassification (mean generalization error, GE) is

\[
E P_n^F \equiv \Phi\{-1/2)\delta_{\text{effective}}\},
\]

where

\[
\delta_{\text{effective}} = \delta/\sqrt{T_n T_x},
\]

is an effective distance between the classes (DeeV, 1970; Raudys, 1972). The term \( T_n = 1 + 2D/(\delta^2 N) \) evaluates an influence of inexactily estimated mean vectors, while the term \( T_x = 1 + D/(n - D) \) evaluates an influence of inexactily estimated covariance matrices.

If \( \bar{N} \to \infty \), then \( E P_n^F \to P_\infty^F = \Phi\{-1/2)\delta\} \), the asymptotic error. In general, the complexity of sample based classification rules depends on a number of parameters estimated from LS and the nature of these parameters (Raudys, 1998b, 2001; Raudys & Young, 2004). For instance, in the two-category case, the dimensionality of the experts’ output vector, \( o = (o_1, o_2, \ldots, o_L)^T \) is \( L \). Then sample estimation of components of an \( L \)-dimensional mean vector \( \mu \) has more influence as an estimation of \( L \) components of variances that are supposed to be the same in two pattern classes (see e.g. Raudys (2001), Raudys and Young (2004)).

The standard deviation of conditional error, \( P_{n_{\Sigma}}^F \) approaches zero if dimensionality, \( L \), and sample size, \( \bar{N} \), are increasing simultaneously and proportionally. In such an asymptote, the sample size/complexity ratio, \( \gamma = \bar{N}/L \), does not change. Both the expected and the conditional PMC, however, are approaching a constant value. This means that in high dimensional cases, regardless of which randomly selected DS is used, approximately the same generalization errors can be obtained each time (Raudys & Jain, 1991).

2.2. Generalization error of the Behavior-Knowledge-Space fusion rule

Theoretical analysis of the BKS rule is of great practical importance since the popular decision tree classifiers are in fact pruned and tailored versions of multinomial classifiers. Consider the two-category case with equal prior probabilities of the classes. There are \( m = 2^L \) potential combinations of the \( L \) experts’ outputs, \( o_1, \ldots, o_L \). To create the ideal fusion rule, the probabilities of all \( 2m \) combinations, \( P_{(1)}^{(s)}, P_{(2)}^{(s)}, \ldots, P_{(m)}^{(s)} \) \((\sum_{r=1}^{m} P_r^{(s)} = 1, s = 1, 2)\), must be known. Then, the Bayes decision rule should make the classification according to the decisions of the \( L \) expert classifiers: allocate the vector \( o \), falling into the \( r \)th cell according to the maximum of \( P_r^{(1)}, P_r^{(2)} \). If \( P_r^{(2)} = P_r^{(1)} \), then ties can be broken randomly.

Denote the number of the \( s \)th learning vectors in the \( r \)th class as \( n_r^{(s)} \). The maximum likelihood estimates of \( P_r^{(1)} = t_r \) and \( P_r^{(2)} = u_r \) are: \( t_r = n_r^{(1)}/N \) and \( u_r = n_r^{(2)}/N \).
correspondingly. If \( \hat{N} \) is increasing without bound, the expected error approaches the Bayes error: \( E P^BKS_n \rightarrow P_B = \frac{1}{2} \sum_{r=1}^{m} \min(\bar{t}_r, u_r) \). The number of training vectors, \( n^{(s)}_i \), in each single cell can be considered as multinomial random variables. After some algebraic manipulation utilizing multinomial distribution properties, the following is obtained:

\[
E P^BKS_n = \frac{1}{2} \left\{ \sum_{r=1}^{m} \left[ t_r P[\hat{t}_r < \hat{u}_r] + u_r P[\hat{t}_r > \hat{u}_r] \right] \right. \\
+ \frac{1}{2} \left[ (t_r + u_r) P[\hat{t}_r = \hat{u}_r] \right] \right\} \\
= \frac{1}{2} \left\{ \sum_{r=1}^{m} \sum_{l=0}^{N-1} \sum_{v=0}^{N-1} \frac{\hat{N}!}{v!(\hat{N} - v)!} u^{t_r}_r (1 - u_r)^{\hat{N} - v} \right. \\
+ \hat{N} \sum_{l=1}^{N-1} \sum_{v=0}^{N-1} \frac{\hat{N}!}{v!(\hat{N} - v)!} u^{t_r}_r (1 - u_r)^{\hat{N} - v} \\
\left. \left. \times \frac{\hat{N}!}{v!(\hat{N} - v)!} u^{t_r}_r (1 - u_r)^{\hat{N} - v} \right] \right\}.
\]

To calculate the generalization error, the \( 2(m - 1) \) probabilities, \( t_r, u_r \), must be known a priori.

2.3. Illustrations of small sample effects

Data used in experiments

To illustrate some of the peculiar characteristics of small-sized learning sets, the two class 8-dimensional (8D) satellite spectral data and multiple classifiers system with seven base experts (MLP with \( h = 30 \) hidden units) were used. The design set, SET1, consists of 4384 and 4242 vectors from the first and second category, respectively. In addition, 3555 and 3606 vectors form the test set, SET2. To design the base classifiers, seven different subsets composed of every 23rd vector of SET1 were selected. Remaining vectors of SET1 encompassed the validation set to be used to determine stopping moment in the Levenberg–Marquardt training procedure. The MLPs were trained five times starting from different initial sets of weights to obtain a high-quality base expert. The seven base experts were made permanent and used in future experiments. After that, the vectors of SET1 and SET2 were classified by the seven experts. A new 7D (7-dimensional) binary valued data, EXP1 (design set) and EXP2 (test set), were used in forthcoming theoretical and experimental studies of the fusion rules.

Fisher classifier based fusion rule

All 15,787 vectors in EXP1 and EXP2 were utilized to estimate the Mahalanobis distance. The theoretical estimate of generalization error (solid curve 1 in Fig. 1) is calculated using Eqs. (1) and (2).

In experimental evaluation of the generalization error, each class in data EXP1 was shuffled at first. Then it was partitioned into 101 non-intersecting subsets \( TS_1, TS_2, \ldots, TS_{101} \). At first, each subset was composed of \( N = 35 \) binary training vectors from each category. The 101 learning subsets were used to perform 101 trials of the experiment. After each single trial with a small learning subset, set EXP2, was used to estimate the generalization errors. We direct readers’ attention to that, in present consideration of the combiners, we made use of different strategies to select vectors to train of the expert classifiers and to perform small sample studies of trainable fusion rules. Afterwards, analysis of fusion rules were repeated with learning set sizes of \( \hat{N} = 38 \) (93 trials with non-intersecting training sets), 43 (82 trials), . . . , 1160 (three trials), 1750 (two trials) and 3500 (one trial). After the first run, EXP1 and EXP2 datasets were interchanged (like in two-fold cross-validation). As shown in Fig. 1, the estimated averages (denoted as “x”) demonstrate that Eq. (1) is accurate even if crisp base expert outputs are used in weighted average fusion.

Fig. 2 shows a scatter diagram of 126 bi-variate vectors \( (P^{(1)}_r, P^{(2)}_r), (r = 2, 3, \ldots, 127) \), the probabilities \( P^{(1)}_r, P^{(2)}_r \) in 126 cells estimated from merged EXP1 and EXP2 data. A box in the upper-right corner shows details of the distribution of smallest values of \( (P^{(1)}_r, P^{(2)}_r) \). Probabilities of the 1st and the 128th cells differ from the remaining ones notably: \( P^{(1)}_1 = 0.3879, P^{(2)}_1 = 0.0125 \) and \( P^{(1)}_{128} = 0.0063, P^{(2)}_{128} = 0.3182 \). So, these cells are not depicted in Fig. 2.

Evaluated values of \( P^{(1)}_1, \ldots, P^{(1)}_{127}, P^{(2)}_1, \ldots, P^{(2)}_{127} \) show that all 128 cells have non-zero probabilities. For the majority of the cells, these probabilities are very small. The Bayes error is \( P_B = 0.0623 \). The probabilities of \( P^{(1)}_r, P^{(2)}_r \) along with Eq. (3) were used to calculate the expected classification error as a function of \( \hat{N} \), the training set size (represented as dotted curve in Fig. 1).

The experimental estimation of the generalization error was carried out in exactly the same way as in the previous trials with
the linear Fisher classifier. The average values of generalization error are depicted as circles in Fig. 1. It can be seen that
theoretical Eq. (3) is nearly accurate.

Comments

A special application of Eq. (3) arises if one employs a
decision tree classifier to fuse the crisp outputs of the experts.
In fact, the decision tree classifier is the pruned version of the
multinomial classifier. A simple mechanistic reduction in the
number of leaves without changing the original decision rule
actually does not change the decision-making rule. Usually
a decision rule with notably smaller number of leaves and
having a larger number of re-substitution errors is simpler than
a decision tree with a smaller number of classification errors.
For that reason, the number of final leaves must be reduced in
the small-sample situation without paying attention to minor
increases in the re-substitution error. Accordingly, Eq. (3) could
be used as a guide in order to determine correct balance between
decision tree complexity and classification error.

3. Complexities of the fusion rules and expert classifiers

3.1. Scissors effect while training the fusion rules

Two learning curves crisscross each other in Fig. 1 and
this resembles a pair of scissors. Fig. 1 advocates that: (1)
in the small sample size situations (where \(N < 1700\)), one
has to use a simple linear weighting fusion rule; (2) in the
large sample size situations (\(N > 1700\)), one may use a more
complex nonlinear multinomial classifier. The “scissors effect”
was described 35 years ago in statistical pattern recognition
publications (Kanal & Chandrasekaran, 1971; Raudys, 1970,
1998a, 2001). Now we demonstrated the scissors effect in the
fusion rule training. Along these lines, for extremely small
sample sizes one needs to use exceptionally uncomplicated
rules, such as simple average or majority voting.

3.2. SA versus WA rule in fusion of continuous outputs of linear experts

Section 3 of the Part I paper presented an intriguing example with continuous expert outputs where a weighted
average fusion rule outperformed a simple average rule ten
times (\(P_{WA}^\infty = 0.0039\) compared to \(P_{SA}^\infty = 0.039\)). In
this data model, asymptotic classification errors of the experts
were equal among themselves. The large differences between
asymptotic errors of WA and SA rules were explained by the
fact that the experts’ outputs were correlated in a special
manner (Eq. (5) in Section 3.1 of Part I). However, large sample
time-based conclusions disagree with numerous observations
by other authors (Sirlantzis, Hoque, & Fairhurst, 2002; Fumera
& Roli, 2005) that claim small differences between the SA and
WA fusion rules. Therefore, the objective of this section is to
elucidate the origin of this disagreement and indicate ways of
knowledgeable MCS designer’s behavior.

To reveal specific relations between the experts’ complexity,
training set size and generalization error, the multivariate
Gaussian data model \(W^k\), discussed in Part I of the paper,
was expanded. We investigate a sequence of data models,
\(W^k\), with increasing input dimensionality, \(p\), where five
base experts are based on different feature subsets. For that
reason, dimensionality, \(k\), used to design each single expert,
is also increasing with an increase in \(p\). So, the input vector
\(x = (x_1, \ldots, x_k, x_{21}, \ldots, x_{2k}, x_{31}, \ldots, x_{3s}, x_{51})\) is separated
into five blocks (sub-vectors), one for each base expert. The
total dimensionality is \(p = 5 \times k\). Correlations between the
experts’ outputs, however, is kept unchanged. The asymptotic
classification error of any single expert does not depend on \(k\),
i.e. \(P_{Exp} = 0.08\).

In the data model \(W^k\), it was assumed that both pattern
classes share a common covariance matrix. The variances of all
\(p\) features is set equal to one. The components \(x_{ij}, x_{2ij}, \ldots, x_{kij}\)
in each single block are mutually independent, whereas the
 correlations between the \(ith\) components \(x_{gi}\) and \(x_{jfi}\) in the
gth and the \(fth\) blocks, is \(\rho_{gf}\). Correlations between features
\(x_{gi}\) and \(x_{jfi}\), \((i \neq f)\) are set equal to zero. Mean value of
the \(k\) category \(k\)-dimensional sub-vector \(x_{jki}, x_{j2i}, \ldots, x_{jk}\) is
\(\mu_{jk} = -(1)^f(1/2)^{g/\sqrt{k}}(1 1 1 \ldots 1 1)\), where \(\delta^1 = 2.81\).

In the MCS analysis, the Fisher classifiers served as the base
experts. The learning set size from one pattern class is fixed
to \(N = 100\) vectors, while dimensionality, \(k\), varied from 5 to
160. Thus, sample size to feature complexity ratio of the
fusion rule, \(y_{fusion} = N/L\), was kept constant. For each single
expert, however, the sample size to feature ratio, \(y_{Expert} = N/k\),
decreases with an increase in \(k\).

One hundred independent experiments using a hundred
\((5k)\)-dimensional Gaussian training sets, \(TS^k\), were performed.
In this experiment, the same data were used to train the
base experts and WA (Fisher DF and Euclidean distance
classifier) fusion rules. For Gaussian data, utilizing linear expert
classifiers with continuous outputs and linear fusion rules the
generalization errors can be calculated analytically. Fig. 3
depict mean values of generalization errors of SA and WA
fusion rules (graphs 1 and 2).

Curves 1 and 2 in Fig. 3 advocates that the scissors effect is also valid in base expert design. For the data model considered,
the Fisher classifier based trainable fusion rule outperforms SA
and EDC based fusion rules if the sample size used to train
each single expert is large (\(N > (1/2)k\), \(y_{Expert} > 1/2\)). With an
increase in input vector dimensionality, \( k \), of each single expert, the performance of a Fisher classifier based trainable fusion rule degrades. When \( k > 2N \), a fixed fusion rule becomes more preferable (the right part of Fig. 3).

Classification performance degradation is caused by two factors:

1. An increase in expected error of each expert, \( E P_{\text{exp}} \), due to an increase in dimensionality, \( k \), (here, \( \delta_{\text{effective}} = \delta((1 + 2k/\bar{N}\delta^2)n/(n-k))^{-1/2} \) in Eq. (2)),

2. Deterioration of the correlation structure between the experts’ outputs.

Examination of sample estimates of correlation matrices has demonstrated that as \( k \) increases, averages of absolute values of the correlations are decreasing rapidly. Using Eq. (3) from Part I, \( E P_{\text{exp}} \) was calculated for different values of \( k \). Suppose for a moment that an ideal correlation structure between the experts’ outputs, \( \Sigma = (\rho_{ij}) \), could be conserved. Then it could be used to find an “imaginative” classification error of MCS,

\[
P_{\text{imaginative}}^{\text{FS}} = \Phi \left\{ -(1/2) \delta_{\text{effective}} \sqrt{\sum_{i=1}^{5} \sum_{j=1}^{5} \sigma_{ij}^{gf}} \right\}^{1/2},
\]

where \( \Sigma^{-1} = (\sigma_{ij}^{gf}) \).

Curve 2a in Fig. 3 confirms that the main increase in generalization error is caused by the distorted structure of correlations between outputs of expert classifiers if \( p < 2N \).

For Gaussian data with a common covariance matrix, the asymptotically optimal is the standard linear Fisher DF. In very small sample situations, the Fisher classifier can perform badly. Usually in such cases, regularized discriminant analysis (RDA) with an optimal regularization parameter is a very good choice. Curves 3a, 3b and 3c in Fig. 3 show the generalization error of RDA in the \( p \)-dimensional feature space designed for three values of the regularization parameter \( \lambda^2 \). Fig. 3 demonstrates that in principle, for certain experts’ learning set size/dimensionality proportions, the optimized RDA can outperform the multiple classifiers system with simple average, EDC and Fisher DF fusion. We remind that early stopping of the single layer perceptron can result in the classifier that is equivalent to RDA (Raudys, 1998a, 2001).

The knowledge that the experts are formed of distinct non-overlapping blocks of input feature vectors is a priori knowledge about the structure of the covariance matrix. This knowledge was used in an inexplicit way when the input feature vector was split into five parts. Curves 3a, 3b and 3c show that if \( 100 < p < 500 \), the MCS could outperform very good classification method, RDA. It is a result of successful utilization of the prior knowledge.

### 3.3. Diversity of the base classifiers and sample size

Diversity of the base classifiers affects selection of the proper fusion rule. In situations where base experts are of similar performance, a simple average often performs better than a weighted average fusion rule. This statement is almost generally admitted and readers are referred to comments in Kuncheva (2004) and Fumera and Roli (2005). Our objective is to clarify this effect theoretically.

To study consequences of the diversity of the base experts on relative performance of WA and SA fusion in small sample size cases we performed experiments with artificial palm data (see Part I). The MLP with seven hidden units were used as the base experts. The MLPs were trained for 10,000 iterations by the standard Levenberg–Marquardt method and 100 independent experiments with different randomly selected subsets of size \( N = 700 \) or 120 were conducted. Two types of seven experts’ aggregates were formed:

(a) Aggregates A (classification errors of single experts varied between 3.1% and 8.2%),

(b) Aggregates B (classification errors varied between 4.1% and 4.5%).

Correlations between outputs of the base networks varied uniformly in an interval \((0.1, 0.7)\).

Weighted average fusion rules were realized by the SLPs. To eliminate effects of imprecision of the fusion rules, WA rules were trained on a very large learning set, \( 1000 + 1000 = 2000 \) of vectors in 2D space. To train fusion rules properly the experts’ output data were decorrelated prior to perceptron training. Then the mean vector of the 7D learning set was calculated in order to shift the data to the center of the coordinates (details of this perceptron training technique can be found in Raudys (2001, chap. 5)). In the SA fusion rule, all weights are equal to 1, i.e. \( g(\omega) = \sum_{i=1}^{5} o_{i} + w_{0} \). Only a threshold term, \( w_{0} \), was adjusted to minimize classification error. Generalization errors were estimated on a test set composed of 1000 vectors of each class. Table 1 shows the average values of the generalization errors obtained in five trials with randomly formed aggregates.

Data in Table 1 confirms the general opinion stating that WA outperforms SA fusion when the experts differ in accuracy.
(set A aggregates). In our example, the differences between trainable and fixed fusion rules are significant. Only in a situation where the variation between the accuracies of the individual experts is minor (aggregates B) and the learning set size is small (case of $N = 120$) does the SA fusion rule produce accuracies comparable to those of the trainable WA fusion. This experiment simply advocates the importance of paying attention to the size of the learning set in addition to the diversity of the base experts. The experimental results are in agreement with the theoretical conclusions obtained in Sections 3.1 and 3.2 where we stated that variability of the base classifiers as well as sample size/complexity ratio used to train both the base classifiers and the fusion rule are important while selecting the architecture of MCS. These newly demonstrated relationships have not yet been considered in the literature.

4. The experts’ self boasting phenomenon

4.1. The experts’ self boasting when Fisher classifier is used as experts

To train the fusion rule in MCS and neural network ensembles design one needs the second independent learning set. In this section we will consider the problems that arise if one does not follow this instruction: training vectors $x^j_l$ ($l = 1, \ldots, N_j; s = 1, 2$) are used to design both the experts and the fusion rules. We will consider the MCS where standard linear Fisher classifiers are used as experts. Theoretical analysis of the Fisher classifier is important since the SLP can realize Fisher DF under certain training conditions.

If the fusion is trained on the same data set as the experts, the fusion rule designer is using the re-substitution error estimates of the DF $g_j(x^j_l) = \hat{w}_j^T x^j_l + \hat{w}_j0$ ($j = 1, 2, \ldots, L$). It is a well-known fact that in the small training set cases, the re-substitution error estimate (empirical or apparent error) is optimistically biased (Amari, 1993; Duda et al., 2000; Fukunaga, 1990; Haykin, 1999; McLachlan, 1992; Raudys, 2001). Evidently, the experts’ self-evaluations, i.e., the re-substitution classification error estimates of the $j$th expert, $E_{\text{re-substitution}}$, are optimistically biased too (Raudys, 2001, Section 6.3.1.2; Janelianus & Raudys, 2002):

$$E_{\text{re-substitution}} \approx (1/2)\delta_{\text{re-substitution}}^j,$$

where $\delta_{\text{re-substitution}}^j = \delta^j(T^j_{\text{re-substitution}})^{1/2}$ and the terms $T_{\text{re-substitution}}$, $T_{\Sigma}$ of Eq. (2) are expert-specific.

In the small sample size and high dimensionality cases, differences between the distances $\delta^j_{\text{effective}} = \delta^j(T^j_{\text{re-substitution}})^{-1/2}$ and $\delta_{\text{re-substitution}}^j = \delta^j(T^j_{\text{re-substitution}})^{1/2}$ can be sizeable. For example, let the dimensionality of input feature set assigned to the expert be $p = 50$, asymptotic error rate $P_{\text{expert}} = 0.2$ ($\delta_{\text{expert}} = 1.68$), the number of training vectors from one pattern class, $N = 250$. Calculations according to Eqs. (1), (2) and (4) result in that expected generalization and re-substitution error rates are similar: $E \hat{P}_{\text{re-substitution}} \approx 0.228$ and $E \hat{P}_{\text{re-substitution}} \approx 0.176$. Let the second expert be much worse, $P_{\text{expert}} = 0.46$ ($\delta_{\text{expert}} = 0.2$), however, it works in high-dimensional space ($p = 250$). Then, $E \hat{P}_{\text{re-substitution}} \approx 0.496$ and $E \hat{P}_{\text{re-substitution}} \approx 0.156$. It means that while training the adaptive fusion rule, the second expert “will say” to the fusion rule designer, that “I perform very well, even better than the first expert”. That’s why, in order to make the fusion rule reliable, instead of the biased estimates of the DF, $g_j(x^j_l)$, the designer needs to use corrected (unbiased) values.

4.2. Reduction of experts’ boasting bias when Fisher classifier is used as the fusion rule

The biased quantity $\delta_{\text{re-substitution}}^j = \delta^j(T^j_{\text{re-substitution}})^{1/2}$ can be expressed as

$$\frac{E_{\text{re-substitution}} - E_{\text{effective}}}{\sqrt{\hat{V}(x^j_l)}}$$

where $E_{\text{re-substitution}}$ denotes the mean of the $g_j(x^j_l)$ if $x^j_l \in \Pi_j$. The term $\sqrt{\hat{V}(x^j_l)}$ denotes the pooled standard deviation of the $g_j(x^j_l)$ and $g_j(x^j_l)$. It was assumed also that $\sqrt{\hat{V}(x^j_l)}$ is common to both categories. The above representation means that, on average, each of the DF values, $g_j(x^j_l)$ and $g_j(x^j_l)$, of categories $\Pi_1$ and $\Pi_2$, respectively, are shifted towards each other by a constant $\Delta g_j$

$$\Delta g_j \equiv (1/2)(\delta_{\text{re-substitution}}^j - \delta_{\text{effective}}^j)\sqrt{\hat{V}(x^j_l)}. \quad (5)$$

Hence, in order to reduce the $j$th expert’s self boasting bias, we recommend to use $\Delta g_j$ in order to shift the discriminant function values of each pattern class

$$\delta_{\text{corrected}}^j(x^j_s) = \hat{w}_j^T x^j_s + \hat{w}_j0 + (-1)^s \Delta g_j,$$

($j = 1, 2, \ldots, L, s = 1, 2$).

Note that the bias value, $\Delta g_j$, is expert-specific. Standard methods of multivariate analysis to estimate parameters of multivariate Gaussian distributions and to calculate the shift term from these estimates are influenced by outliers, atypical observations. To obtain more robust estimates, firstly we recommend to estimate the re-substitution errors $\hat{P}_{\text{re-substitution}}$. Once the re-substitution error $\hat{P}_{\text{re-substitution}}$ is estimated, the unknown $\delta_{\text{re-substitution}}^j$ and the Mahalanobis distance $\delta^j$ can be estimated by means of computer-based interpolation of these two equations:

$$\hat{P}_{\text{re-substitution}} = \Phi(-1/2)\delta_{\text{re-substitution}}^j \quad \text{and} \quad \delta_{\text{re-substitution}}^j = \delta^j(T^j_{\text{re-substitution}})^{1/2}$$

To evaluate the effective distance $\delta_{\text{effective}}^j$, one needs to use $\delta^j$ and Eq. (2). Standard deviation, $\sqrt{\hat{V}(x^j_l)}$ in Eq. (5) can be estimated as the ratio of the squared difference in mean values, $E_{\text{re-substitution}}(x^j_l) - E_{\text{re-substitution}}(x^j_l)$, and the term $\delta_{\text{re-substitution}}^j$.

<table>
<thead>
<tr>
<th>$N$</th>
<th>Set A aggregates</th>
<th>Set B aggregates</th>
</tr>
</thead>
<tbody>
<tr>
<td>700</td>
<td>0.022</td>
<td>0.045</td>
</tr>
<tr>
<td>120</td>
<td>0.025</td>
<td>0.063</td>
</tr>
</tbody>
</table>

Table 1
Generalization errors of the weighted average and simple average fusion rules

<table>
<thead>
<tr>
<th>$N$</th>
<th>WA</th>
<th>SA</th>
<th>WA</th>
<th>SA</th>
</tr>
</thead>
<tbody>
<tr>
<td>700</td>
<td>0.022</td>
<td>0.045</td>
<td>0.014</td>
<td>0.037</td>
</tr>
<tr>
<td>120</td>
<td>0.025</td>
<td>0.063</td>
<td>0.032</td>
<td>0.030</td>
</tr>
</tbody>
</table>
The experts’ bias-reduction technique was based on analytical expressions of the mean values of the classification errors. It ignores variances caused by the random character of learning sets. For that reason, the procedure suggested will be effective if the variances of the re-substitution estimates \( \hat{P}_{Rj} = n_{\text{errors}}/n \) (\( j = 1, 2, \ldots, L \)) and that of the conditional classification errors \( P_{Rj} \) are small.

There exist situations, however, where both the number of dimensions and the sample size are large. Then standard deviation of the conditional classification error of the Fisher DF, \( \sqrt{V(P_{Rn})} \), is small compared to an increase in generalization error \( E(P_{Rn}) - P_{R}^F \). Irrespective of which randomly selected LS is used in the high-dimensional problems, one will obtain approximately the same conditional error. This state of affairs is similar to the situation with standard deviation of the re-substitution (apparent) classification error estimate \( \hat{P}_R \). Independent of the dimensionality, the standard deviation of estimate \( \hat{P}_R \) is given by Eq. (7).

\[
\left( V \hat{P}_R \right)^{1/2} \approx \left( E \hat{P}_R \left( 1 - E \hat{P}_R \right)^{1/2} \right) \frac{1}{n}.
\]

Standard deviation tends to zero if the learning set size, \( n \), is increasing without bound. Thus, in high dimensional large sample size tasks, the experts’ bias reduction technique could be effective. For the bias reduction technique to be highly efficient, another condition has to be satisfied, that is: the classification errors, dimensionalities and training set sizes of the individual expert classifiers should be different. Otherwise, the modification terms \( \Delta g_j \) in Eq. (6) will be identical, and hence ineffective.

**Experiments**

In order to illustrate the effectiveness of the bias reduction technique and compare it with traditional fusion rules, simulation experiments with four specially designed artificial datasets were performed. Seven linear Fisher DF based experts were used in all four MCS investigated.

A Gaussian dataset \( A1 \) was designed to demonstrate the usefulness of the expert bias elimination technique: dimensionality and discriminative power of input feature subsets assigned to distinct experts were different. The data composed of 1350 independent Gaussian features were split among seven base classifiers, each having 250, 250, 250, 250, 50 or 50 features. Asymptotic errors for the first five among seven base classifiers, each having 250, 250, 250, 250, 1350-dimensional space was set to \( \frac{1}{2} \). Asymptotic errors for the first five among seven base classifiers, each having 250, 250, 250, 250, 1350-dimensional space was set to \( \frac{1}{2} \).

The Gaussian dataset \( A2 \) was designed to demonstrate the ineffectiveness of expert bias elimination technique: dimensionality and discriminative power of input feature subsets were the same for all experts. 1150 independent Gaussian input features were divided among the base classifiers with an overlap of 100 features: 1–250; 151–400; 301–550; 451–700; 601–850; 751–1000 and 901–1150 (250 features for each expert). All experts were equally qualified. The asymptotic errors of the experts is \( P_{\text{expert}} = 0.2387 \). Asymptotic performance of Fisher classifier in 1150D feature space, \( P_{\infty}^F = 0.03 \).

The Gaussian data \( B \) represents a multivariate data model with different covariance matrices, a situation where both the experts and trainable fusion rule (the Fisher classifier) turn out to be asymptotically non-optimal decision-making rules. The data \( B \) has 1750 input features, 250 features for each expert. The features within each of the seven groups were statistically independent. Only a few correlations between outputs of several expert classifiers were relatively high. In the first class, \( \rho_{14} \approx 0.94; \rho_{17} \approx 0.61; \rho_{25} \approx 0.85; \rho_{36} \approx 0.54; \rho_{47} \approx 0.64 \), while the remaining correlations were close to zero. In the second class, signs of correlation coefficients were opposite.

20-dimensional and 3-dimensional datasets \( V_v \) were designed to demonstrate usefulness of the expert bias elimination technique for BKS fusion rule to be discussed below. This data model represents situations where the Fisher classifiers (SLPs as well) are unsuitable for classification purposes. The first two components of each pattern class were mixtures of two bi-variate Gaussian components (subclasses). The distribution densities of each of the categories is reminiscent of the letter “V” in the 2D space. Vectors of the first class are “situated inside” the vectors of the second class. Variances of the 2D Gaussian subclasses were equal to one. Both features were highly correlated (correlation coefficient \( \rho = 0.999 \)). For that reason, an overlap of patterns vectors that belong to distinct classes is negligible. The other \( p-2 \) features are uncorrelated uninformative Gaussian \( N(0, 0.5^2) \) random variables. Accordingly, the Bayes error exceeds 0.03 only insignificantly. The experts were trained on different, although overlapping parts of the design set. In this split of design set, only two out of seven experts were trained with uni-modal Gaussian data. Only these two experts were good authorities in their own areas of the feature space. The other five experts were trained with Gaussian mixtures and their performances were bad.

Results obtained in 25 trials with independent design sets are presented in Table 2. The upper and lower cell values represent averages and standard deviations, respectively. The first row in Table 2 is obtained for dataset \( A1 \) and relatively small training set sizes (\( \bar{N} = 250 \), sample size to feature ratio, \( \gamma_{\text{expert}} = N/p = 250/1350 \approx 0.1852 \)). The first row indicates that the bias reduction technique diminishes the generalization error of the standard Fisher fusion rule by a factor of three (from 0.2216 to 0.0749).

For large training set size (\( \bar{N} = 2500 \)), however, both the error rates and the gain are smaller. Using standard Fisher fusion the generalization error is \( P_{\text{gen}} = 0.0361 \) while for the Fisher rule with bias reduction, the error is only a little bit smaller, \( P_{\text{gen}} = 0.0329 \). The SA fusion rule gave \( P_{\text{gen}} = 0.0361 \). Due to rather different classification performances of the experts, the majority voting fusion rule resulted in much worse classification, \( P_{\text{gen}} = 0.2227 \).

In datasets \( A2 \) and \( B \), all features were equally relevant and all seven experts were working in 250-dimensional feature subspaces. The bias terms, \( \Delta g_1, \Delta g_1, \ldots, \Delta g_7 \) (Eq. (5)), were...
almost equal among themselves and actually unnecessary. Consequently, the bias reduction technique was practically useless. The modified Fisher fusion rule failed to outperform the standard Fisher rule. Efficacy of both SA and WA fusion rules are almost similar. Analogous conclusions can be drawn from experiments using data \( \mathbf{V} \).

### 4.3. Reduction of experts’ boasting bias when BKS strategy is used as the fusion rule

Consider MCS where the Fisher linear classifiers are used as the base experts and crisp outputs are submitted to the BKS fusion rule. Like in the previous subsection, a learning set is used to train both, the experts and the fusion rule. Obviously, cell probabilities \( \hat{P}^{(1)}_1, \hat{P}^{(1)}_2, \ldots, \hat{P}^{(1)}_{n-1} \) of the multinomial classifier will be biased. In order to reduce the bias it was suggested to use the experts’ independence assumption and unbiased estimates of the generalization error of single experts (Raudys, 2003).

The previous subsection’s way utilized to reduce the expert bias is sensitive to outliers, atypical observations in the training sets. It could become a problem if one deals with non-Gaussian data. Like in the previous subsection, in order to get more robust estimates of generalization error, the Mahalanobis distance \( \delta^{(j)} \) specific to the \( j \)th expert could be estimated from re-substitution error \( \hat{P}^{(s)}_{R,j} \). In this approach, distance \( \delta^{(j)} \) is evaluated in two stages. In the first stage, one interpolates equation \( \hat{P}^{(s)}_{R,j} = \Phi[-(1/2)\delta^{(j)}_{\text{re-substitution}}] \) and from re-substitution error estimate, \( \hat{P}^{(s)}_{R,j} \), finds \( \delta^{(j)}_{\text{re-substitution}} \). Then one interpolates the equation \( \delta^{(j)}_{\text{re-substitution}} = \delta^{(j)}_{\mu G} \mu_{G}^{-1/2} \) and seeks for the estimate \( \hat{\delta}^{(j)} \). Finally, from (1), (2) the GE of the \( j \)th expert is re-evaluated:

\[
\hat{P}^{(s)}_{\hat{\mu}_j} = \Phi[-(1/2)\hat{\delta}^{(j)}(1 + 2p/(\hat{\delta}^{(j)}\sqrt{\hat{N}}))^{1/2} \times (1 + p/(n - p))^{-1/2}].
\]

To evaluate the bias of the maximum likelihood estimate, \( \hat{P}^{(s)}_{(r)} = n^{(s)}_r / N, \) of the \( r \)th cell we suppose that the expert outputs are independent. Then the bias will be estimated as a difference between two product terms

\[
\hat{B}^{(s)}_{r} = \prod_{j=1}^{L} \left( \hat{P}^{(s)}_{R,j} \right)^{2-o_j}(1 - \hat{P}^{(s)}_{R,j} o_j)^{-1} - \prod_{j=1}^{L} \left( \hat{P}^{(F)}_{R,j} \right)^{2-o_j}(1 - \hat{P}^{(F)}_{R,j} o_j)^{-1},
\]

where \( \hat{P}^{(s)}_{R,j} \) and \( \hat{P}^{(F)}_{R,j} \) are the standard re-substitution and generalization error estimates for the \( j \)th expert defined in Eq. (8).

Raudys (2003) suggested adding this term to the maximum likelihood estimate \( \hat{P}^{(s)}_{(r)} \). The rightmost column of Table 2 indicates that this idea is efficient for statistically independent and dependent experts, provided that: (i) the input data is unimodal, (ii) the experts are of different quality and dissimilar complexity, and (iii) the training set size/complexity ratio is relatively small (data sets \( \text{A1, A2} \) for \( \hat{N} = 250 \) and data set \( \text{B} \)). This technique is invalid if the aforementioned conditions are violated: i.e., the performance and complexity of the experts are similar, the training set sizes are sufficiently large, and/or the dataset is multi-modal.

### 4.4. A noise injection technique to deal with the boosting bias and complexity problems

The bias correcting terms of Eqs. (5) and (9) may be useful if theoretically determined assumptions can be satisfied. In sum, the results presented in the previous two subsections demonstrate that principal possibility exists to improve performance of MCS in finite sample size situations. Unfortunately, these theory-based results cannot be exploited always. If an independent validation set is unavailable, Wolpert (1992) suggested on using leave-one-out estimates of the experts’ outputs to overcome the boosting bias.
Another possible way to deal with the small sample size problem is to form pseudo-training or pseudo-validation sets by means of a noise injection. This can be accomplished by adding random zero mean vectors to each training vector. In Raudys (2003), a colored k-NN noise injection was used to form artificial pseudo-validation sets. Then one classifies vectors of the pseudo-validation set by the base experts, forms a new L-dimensional pseudo-artificial learning set and uses it to find parameters of the fusion rule.

To generate colored noise, for each single training vector, \( x^{(s)}_1 \), one finds its \( k \) nearest neighbors of the same pattern class and adds a noise only in a subspace formed by the vector \( x \) and the \( k \) neighboring training vectors, \( x^{(s)}_1, x^{(s)}_2, \ldots, x^{(s)}_k \) (Duin, 1993; Skurichina, Raudys, & Duin, 2000; Raudys, 2001). Random Gaussian \( N(0, \sigma^2_{\text{noise}}) \) noise is added \( n_{\text{nn}} \) times along the \( k \) lines connecting \( x^{(s)}_1 \) and \( x^{(s)}_1, x^{(s)}_2, \ldots, x^{(s)}_k \). Three parameters are required to realize the noise injection procedure: \( k \), the number of neighbors, \( n_{\text{nn}} \), the number of new, artificial vectors generated around a single training vector, \( x^{(s)}_1 \), and \( \sigma_{\text{noise}} \), the noise standard deviation. The noise variance, \( \bar{\sigma}^2_{\text{noise}} \), has to be selected as a trade-off between the complexity of the decision boundary and the LS size. When working with unknown data, one has to test several values of \( \sigma_{\text{noise}} \) and select the most suitable one.

In Fig. 4, we show four graphs of the generalization error obtained in the experiments with BKS fusion rule and the artificial \( \mathbf{Vv} \), 20D \((n = 60 + 60)\) and 3D \((n = 12 + 12)\) data. The graphs correspond to mean values of 50 independent trials with diverse randomly selected training sets. To train the experts, the learning set was used. To train the BKS fusion rule, the \( L \)-dimensional pseudo-artificial training set was used. The training set sizes were “increased” \( n_{\text{nn}} \)-fold by adding 2-NN vectors \((n_{\text{nn}} = 4 \text{ or } 20)\) to the original vectors \( x^{(1)}, x^{(2)}, \ldots, x^{(N)} \). All four graphs in Fig. 4 exhibit clear minima. In the minima region, for \( p = 20, \bar{N} = 60 \) the generalization error is \( P_{\text{gen}} \approx 0.105 \) while for \( p = 3 \) and \( \bar{N} = 12 \), \( P_{\text{gen}} \approx 0.12 \). It is notably smaller compared to the generalization error of the standard BKS rule \((0.1478 \text{ and } 0.2023)\) or its theoretical modification \((0.12)\).

In short, it can be concluded that the noise injection technique has a double effect:

1. It reduces the fusion rule’s complexity and
2. It diminishes the expert’s boasting bias.

The noise injection approach, however, cannot create entirely new information and improve everything. If DS is very small, the data points incorrectly represent the general population. The noise injection cannot restore unknown true data. So, at times it becomes useless. The noise injection with incorrect values of noise variance can become even harmful. In the high-dimensional space, one has to use larger \( n_{\text{nn}} \) values in order to fill the empty spaces between training vectors of the same pattern class. As a result, when working with extremely high-dimensional data, the application of the noise injection technique requires large computer resources and becomes ineffective.

5. Concluding remarks

The paper is focused on the multiple classifiers system and neural network ensemble design from the generalization error point of view. Information used to design a recognition system is composed of learning and validation data. In neural network ensembles and multiple classifiers system design, one needs to split the learning set in two parts. If one avoids the data partition, he/she faces extra small sample effects. Formal and informal assumptions, certain guesses about the data and its structure could be useful in order to develop techniques aimed to reduce undesirable small sample effects. Several of them have been considered in this paper. We made efforts to elucidate theoretically and demonstrate experimentally in which way the effectiveness of such techniques depends on a degree of violation of the assumptions. We demonstrated that wrong assumptions destroy good methods at times.

In neural network ensembles, one needs to think about: which information (guesses) one has to introduce while dividing the complex pattern recognition task among the expert neural networks and which type of the fusion rule to use. Experimental analyses show that “blindly” splitting the task into subtasks, one for each expert, sometimes becomes useless.

\[
\begin{align*}
\text{Fig. 4. The mean generalization error as a function of noise variance, in the experiments with artificial } \mathbf{Vv} \text{ data: (a) } & p = 20, \bar{N} = 60; \text{ (b) } p = 3, \bar{N} = 12; \text{ (c) } n_{\text{nn}} = 5; \text{ (d) } n_{\text{nn}} = 20. \\
\end{align*}
\]
Nevertheless, one assumption that always appears fruitful is that while splitting complex task into subtasks, we improve sample size/complexity ratio.

In MCS design, ideal fusion of expert decisions leads to a classification system with better performance in comparison with that of the best expert. If a defective fusion rule is used, performance of MCS could even worsen in comparison with the best expert. In MCS, a priori information is available: “positive” outputs of expert classifiers are associated with certain pattern classes. In such situations, non-trainable fusion rules such as simple average, product, majority voting etc. appear to be useful. In the present paper we paid particular attention to two types of trainable fusion rules: the non-linear Behaviour-Knowledge-Space method realized by multinomial or decision tree classifiers and weighted average fusion realized by a single layer perceptron or linear Fisher classifier. If correctly used, both rules become effective tools in the neural network ensembles’ design.

Like in standard pattern recognition tasks, the scissors phenomenon occurring in neural network ensembles suggests that in small sample situations one needs to use simple base experts, the fixed fusion rules or uncomplicated trainable combiners based on small number of inputs. Another specific effect mentioned above, but not yet considered in the literature, is that a small training set reduces the quality of the expert classifiers and destroys the correlation structure of the experts’ outputs. If the training set size, used to develop the expert classifiers is too small, the usefulness of utilizing trainable fusion rules instead of fixed ones becomes problematic.

As for MCS design, the fact that only a fraction of the expert networks was selected suggests that some weighting of the experts has been done already. The selected fraction of the experts was assigned rank “one” while the rejected fraction obtained rank “zero”. It is an introduction of certain prior experts was assigned rank “one” while the rejected fraction has been done already. The selected fraction of the experts has been done already. The selected fraction of the experts was assigned rank “one” while the rejected fraction obtained rank “zero”. It is an introduction of certain prior

in such situations, non-trainable fusion rules such as simple average, product, majority voting etc. appear to be useful. In the present paper we paid particular attention to two types of trainable fusion rules: the non-linear Behaviour-Knowledge-Space method realized by multinomial or decision tree classifiers and weighted average fusion realized by a single layer perceptron or linear Fisher classifier. If correctly used, both rules become effective tools in the neural network ensembles’ design.

Like in standard pattern recognition tasks, the scissors phenomenon occurring in neural network ensembles suggests that in small sample situations one needs to use simple base experts, the fixed fusion rules or uncomplicated trainable combiners based on small number of inputs. Another specific effect mentioned above, but not yet considered in the literature, is that a small training set reduces the quality of the expert classifiers and destroys the correlation structure of the experts’ outputs. If the training set size, used to develop the expert classifiers is too small, the usefulness of utilizing trainable fusion rules instead of fixed ones becomes problematic.

As for MCS design, the fact that only a fraction of the expert networks was selected suggests that some weighting of the experts has been done already. The selected fraction of the experts was assigned rank “one” while the rejected fraction obtained rank “zero”. It is an introduction of certain prior

in such situations, non-trainable fusion rules such as simple average, product, majority voting etc. appear to be useful. In the present paper we paid particular attention to two types of trainable fusion rules: the non-linear Behaviour-Knowledge-Space method realized by multinomial or decision tree classifiers and weighted average fusion realized by a single layer perceptron or linear Fisher classifier. If correctly used, both rules become effective tools in the neural network ensembles’ design.

Like in standard pattern recognition tasks, the scissors phenomenon occurring in neural network ensembles suggests that in small sample situations one needs to use simple base experts, the fixed fusion rules or uncomplicated trainable combiners based on small number of inputs. Another specific effect mentioned above, but not yet considered in the literature, is that a small training set reduces the quality of the expert classifiers and destroys the correlation structure of the experts’ outputs. If the training set size, used to develop the expert classifiers is too small, the usefulness of utilizing trainable fusion rules instead of fixed ones becomes problematic.

As for MCS design, the fact that only a fraction of the expert networks was selected suggests that some weighting of the experts has been done already. The selected fraction of the experts was assigned rank “one” while the rejected fraction obtained rank “zero”. It is an introduction of certain prior

in such situations, non-trainable fusion rules such as simple average, product, majority voting etc. appear to be useful. In the present paper we paid particular attention to two types of trainable fusion rules: the non-linear Behaviour-Knowledge-Space method realized by multinomial or decision tree classifiers and weighted average fusion realized by a single layer perceptron or linear Fisher classifier. If correctly used, both rules become effective tools in the neural network ensembles’ design.

Like in standard pattern recognition tasks, the scissors phenomenon occurring in neural network ensembles suggests that in small sample situations one needs to use simple base experts, the fixed fusion rules or uncomplicated trainable combiners based on small number of inputs. Another specific effect mentioned above, but not yet considered in the literature, is that a small training set reduces the quality of the expert classifiers and destroys the correlation structure of the experts’ outputs. If the training set size, used to develop the expert classifiers is too small, the usefulness of utilizing trainable fusion rules instead of fixed ones becomes problematic.

As for MCS design, the fact that only a fraction of the expert networks was selected suggests that some weighing of the experts has been done already. The selected fraction of the experts was assigned rank “one” while the rejected fraction obtained rank “zero”. It is an introduction of certain prior information expressed in an implicit way. Hence, one may say that if one decides to use the maximum, simple average or majority voting fusion rules, actually he/she already uses a trainable fusion rule with binary (0, 1) weights. Consequently, designing and testing one more, “really trainable”, fusion rule becomes less effective.

In designing MCS with trainable fusion rules, the outputs of expert classifiers become biased if learning data is used to train the fusion rule. The presence of boasting bias introduces serious difficulties. A standard way to overcome the bias effects is to split the learning set into two parts. One is used to train the expert classifiers and the other to train the fusion rule. We have demonstrated that for known data models one may develop theoretical correction terms to reduce the bias and avoid splitting the learning set into two parts. Utilization of analytical correction terms helps to improve MCS’s performance. Such terms have been developed when linear Fisher DF or BKS rules were used for fusion. If assumptions about data models are not satisfied, the correction terms become useless or could even worsen MCS.

A truthful assumption about the data is the compactness hypothesis that appeared in the early days of research in pattern recognition. It states that objects of one pattern class are situated in neighboring areas of multivariate feature space. Modification of this hypothesis leads to a colored noise injection (k-nearest-neighbor-directed noise injection). In this method, in order to minimize distortion of the data, one assumes inexplicitly that local subspace between neighboring training vectors is not empty. Then one has to inject a noise in the subspace of k-nearest vectors of each single training vector. Standard cost functions used in neural network design do not utilize such information. In this analysis, it was shown that exploitation of the k-NN directed noise injection technique in order to form the pseudo-training set aimed to train the fusion rule could be useful.

Neither feature extraction nor neural network ensembles are panaceas that generate information automatically. Therefore, in seeking for ways to improve the performance of neural network ensembles, the designer must make special efforts to examine what particular implicit information is present in his system design strategy.

Acknowledgments

The author thanks Dr. Krystyna Stankiewicz for providing satellite data for experiments, Professors Ray Somorjai, Fabio Roli and Aini Hussain for useful remarks, challenging discussions and aid in preparing the final version of the paper. The author appreciates notable contributions of the anonymous reviewers.

References


