Feature Subsets for Classifier Combination: An Enumerative Experiment

L.I. Kuncheva and C.J. Whitaker
School of Informatics, University of Wales, Bangor
Bangor, Gwynedd, LL57 1UT, United Kingdom
{l.i.kuncheva,c.j.whitaker}@bangor.ac.uk

Abstract

A classifier team is used in preference to a single classifier in the expectation it will be more accurate. Here we study the potential for improvement in classifier teams designed by the feature subspace method: the set of features is partitioned and each subset is used by one classifier in the team. All partitions of a set of 10 features into 3 subsets containing $\langle 4, 4, 2\rangle$ features and $\langle 4, 3, 3\rangle$ features, are enumerated and nine combination schemes are applied on the three classifiers. We look at the distribution and the extremes of the improvement (or failure); the chances of the team outperforming the single best classifier if the feature space is partitioned at random; the relationship between the spread of the individual classifier accuracy and the team accuracy; and the combination schemes performance.

1 Introduction

We examine by an enumerative experiment what the support is for the intuition that a team of classifiers performs better than the single best classifier in the team. The feature subspace method has been used: we partition the set of features into subsets where each subset is used by one classifier in the team. Using different feature subsets has been recognized as a promising team design method, especially in text recognition [14, 17] and speech recognition [2]. Kittler et al. [7, 8] derive a series of theoretical results based on the assumption that the individual classifiers use conditionally independent subsets of features. Sometimes the features are naturally grouped and this suggests which of them should be used together. For example, Duin and coauthors [4] (and earlier [15]) study classifier fusion methods for recognizing handwritten numerals by using 6 types of different features sets: Fourier coefficients, profile correlation, Karhunen-Loève coefficients, pixel averages in 2×3 windows, Zernike moments and morphological features. Random sampling from the feature set for designing the individual classifiers has been studied in [3, 5, 13]. A genetic algorithm for partitioning the feature space is proposed in [9, 11].

Here we offer an exhaustive experimental study with $L = 3$ classifiers and a data set with $n = 10$ features enumerating all partitions of the feature set into $\langle 4, 4, 2\rangle$ and $\langle 4, 3, 3\rangle$ features. Let P_t be the accuracy of the team, P_b be the best (maximal) individual accuracy, and P_w be the worst (minimal) individual accuracy.

We seek answers to the following questions:

1. How is $P_t - P_b$ distributed and what are the maximal and the minimal possible values for different combination schemes?

2. How likely is an improvement ($P_t - Pb > 0$) if we pick a random partition of the set of features?

3. Is the team accuracy P_t related to the range $P_b - P_w$ of individual accuracies?

4. How do the combination schemes compare with respect to the answers to the previous three questions?

Section 2 details the combination methods used, so that they be reproducible from the text. Section 3 contains the results of our experiment and the conclusion section offers the answers to the above questions.

2 Combination methods

Let $\mathcal{D} = \{D_1, D_2, \ldots, D_L\}$ be a set of classifiers and $\Omega = \{\omega_1, \ldots, \omega_c\}$ be a set of class labels. Each classifier gets as its input a feature vector $\mathbf{x} \in \Re^n$. The classifier output is a c-dimensional vector

$D_i(\mathbf{x}) = [d_{i,1}(\mathbf{x}), \ldots, d_{i,c}(\mathbf{x})]^T$ where $d_{i,j}(\mathbf{x})$ is the degree of "support" given by classifier D_i to the hypothesis that $\mathbf{x}$ comes from class ω_j, $j = 1, \ldots, c$. Without loss of generality we can restrict $d_{i,j}(\mathbf{x})$ within the interval $[0, 1]$, $i = 1, \ldots, L$, $j = 1, \ldots, c$, and call the classifier outputs "soft labels" (see [1]). Most often $d_{i,j}(\mathbf{x})$ is an estimate of the posterior probability $P(\omega_i|\mathbf{x})$.

Combining classifiers means we combine the L classifier outputs $D_1(\mathbf{x}), \ldots, D_L(\mathbf{x})$ to get a soft label for $\mathbf{x}$, denoted $D(\mathbf{x}) = [\mu_1(\mathbf{x}), \ldots, \mu_c(\mathbf{x})]^T$.

If a crisp class label of $\mathbf{x}$ is needed, we can use the maximum membership rule: Assign $\mathbf{x}$ to class ω_s iff,

$$\begin{aligned} d_{i,s}(\mathbf{x}) &\geq d_{i,j}(\mathbf{x}) \;\; \forall j = 1, \ldots, c. \;\; \text{for individual crisp labels} \\ \mu_s(\mathbf{x}) &\geq \mu_t(\mathbf{x}), \;\; \forall t = 1, \ldots, c. \;\;\; \text{for the final crisp label.} \end{aligned} \tag{1}$$

Ties are resolved arbitrarily. The minimum-error classifier is recovered from (1) when $\mu_i(\mathbf{x}) = P(\omega_i|\mathbf{x})$.

2.1 Majority vote, Maximum, Minimum, Average, Product

For the majority vote combination (MAJ), the class label assigned to $\mathbf{x}$ is the one that is most represented in the set of L class labels $D_1(\mathbf{x}), \ldots, D_L(\mathbf{x})$. For the remaining simple combination methods,

$$\mu_j(\mathbf{x}) = \mathcal{O}\left(d_{1,j}(\mathbf{x}), \ldots, d_{L,j}(\mathbf{x})\right), \;\; j = 1, \ldots, c. \tag{2}$$

where $\mathcal{O}$ is the respective operation (maximum (MAX), minimum (MIN), average (AVR) or product (PRO)).

2.2 Naive Bayes (NB)

This scheme assumes that the classifiers are mutually independent (this is the reason we use the name "naive"); Xu et al. [17] and others call it *Bayes* combination. For each classifier D_i, a $c \times c$ confusion matrix CM^i is calculated by applying D_i to the training data set. The (k, s)th entry of this matrix, $cm^i_{k,s}$ is the number of elements of the data set whose true class label was ω_k, and were assigned by D_i to class ω_s. By $cm^i_{\cdot,s}$ we denote the total number of elements labeled by D_i into class ω_s (the sum of the sth column of CM^i). Using $cm^i_{\cdot,s}$, a $c \times c$ label matrix LM^i is computed, whose (k,s)th entry $lm^i_{k,s}$ is an estimate of the probability that the true label is ω_k given that D_i assigns crisp class label s.

$$lm^i_{k,s} = \hat{P}\left(\omega_k | D_i(\mathbf{x}) = \omega_s\right) = \frac{cm^i_{k,s}}{cm^i_{\cdot,s}}, \tag{3}$$

Considering the label matrix for D_i, LM^i, associated with ω_s is a *soft label vector* $[\hat{P}\left(\omega_1 | D_i(\mathbf{x}) = \omega_s\right), \ldots, \hat{P}\left(\omega_c | D_i(\mathbf{x}) = \omega_s\right)]^T$, which is the sth column of the matrix. Let $s_1, \ldots, s_L$ be the crisp class labels assigned to $\mathbf{x}$ by classifiers $D_1, \ldots, D_L$, respectively. Then, by the independence assumption, the estimate of the probability that the true class label is ω_j, is calculated by

$$\mu_j(\mathbf{x}) = \prod_{j=1}^{L} \hat{P}\left(\omega_j | D_i(\mathbf{x}) = s_i\right) = \prod_{i=1}^{L} lm^i_{j,s_i}, \quad j = 1, \ldots, c. \tag{4}$$

2.3 Behavior-knowledge space (BKS)

Let again $(s_1, \ldots, s_L) \in \Omega^L$ be the crisp class labels assigned to $\mathbf{x}$ by classifiers $D_1, \ldots, D_L$, respectively. Every possible combination of class labels is an index regarded as a cell in a look-up table (BKS table)[6]. The table is designed using a labeled data set $\mathbf{Z}$. Each $\mathbf{z}_j \in \mathbf{Z}$ is placed in the cell indexed by $D_1(\mathbf{z}_j), \ldots, D_L(\mathbf{z}_j)$. The number of elements in each cell are tallied and the most representative class label is selected for this cell. Ties are resolved arbitrarily and the empty cells are labeled appropriately (e.g., at random or by majority, if applicable). After the table has been designed, the BKS method labels an $\mathbf{x} \in \Re^n$ to the class of the cell indexed by $D_1(\mathbf{x}), \ldots, D_L(\mathbf{x})$.

2.4 Wernecke's method (WER)

The model is similar to the BKS. The difference is that in constructing the table, Wernecke [16] considers the 95 % confidence intervals of the frequencies in each cell. If there is overlap between the intervals, the L confusion matrices are used to identify the "least wrong" classifier among the L members of the team. First, L estimates of the probability $P(error \text{ and } D_i(\mathbf{x}) = s_i)$ are calculated.

Then the classifier with the smallest probability is nominated for labeling the cell. For an $\mathbf{x} \in \Re^n$, the cell is identified by the labels assigned by $D_1, \ldots, D_L$ and then either the cell label is recovered or the label of the nominated classifier is taken as the label of $\mathbf{x}$.

2.5 Decision templates (DT)

The classifier outputs can be conveniently organized in a **decision profile** as the following matrix [10]

$$DP(\mathbf{x}) = \begin{bmatrix} d_{1,1}(\mathbf{x}) & ... & d_{1,j}(\mathbf{x}) & ... & d_{1,c}(\mathbf{x}) \\ \cdots & & & & \\ d_{i,1}(\mathbf{x}) & ... & d_{i,j}(\mathbf{x}) & ... & d_{i,c}(\mathbf{x}) \\ \cdots & & & & \\ d_{L,1}(\mathbf{x}) & ... & d_{L,j}(\mathbf{x}) & ... & d_{L,c}(\mathbf{x}) \end{bmatrix}. \tag{5}$$

Using decision templates (DT) for combining classifiers is proposed in [10]. Given L (trained) classifiers in $\mathcal{D}$, c decision templates are calculated from the data, one per class.

$$DT_i = \frac{1}{N_i} \sum_{\substack{\mathbf{z}_j \in \omega_i \\ \mathbf{z}_j \in \mathbf{Z}}} DP(\mathbf{z}_j), \quad i = 1, \ldots, c. \tag{6}$$

DT_i can be regarded as the expected $DP(\mathbf{x})$ for class ω_i. The support for the class offered by the combination of the L classifiers, $\mu_i(\mathbf{x})$ is then found using a measure of *similarity* between the current $DP(\mathbf{x})$ and DT_i, e.g.,

$$d_E(DP(\mathbf{x}), DT_i) = \sum_{j=1}^{c} \sum_{k=1}^{L} (d_{k,j}(\mathbf{x}) - dt_i(k,j))^2, \tag{7}$$

where $dt_i(k,j)$ is the k,j-th entry in decision template DT_i. Here we use Euclidean distance for calculating the similarity but other measures can also be applied.

3 The experiment

We used the Wisconsin Diagnostic Breast Cancer data base[1] taken from the UCI Repository of Machine Learning Database[2]. The set consists of 569 patient vectors with features computed from a digitized image of a fine needle aspirate of a breast mass. They describe characteristics of the cell nuclei present in the image. The objects are grouped into two classes: benign and malignant. Out of the original 30 features we used the first 10; these were the means of the relevant variables calculated in the image. The study was confined to 10 variables for two reasons: to enable a reasonable enumerative experiment and to enhance variability in classifier performance. The data set was split randomly into two halves, one being used for training and one for testing.

We considered $L = 3$ classifiers. All partitions of the 10-element feature set into $\langle 4,4,2 \rangle$ (3150 partitions) and $\langle 4,3,3 \rangle$ (4200 partitions) were generated. For each partition, three classifiers were built, one on each subset of features. Two simple classifier models were tried: the linear and the quadratic classifier, leading to 4 sets of experiments:

1. $\langle 4,4,2 \rangle$ with linear classifiers;
2. $\langle 4,4,2 \rangle$ with quadratic classifiers;
3. $\langle 4,3,3 \rangle$ with linear classifiers;
4. $\langle 4,3,3 \rangle$ with quadratic classifiers.

To answer the four questions in the Introduction,

1. The minimal and the maximal values of the differences between the accuracy of the team and the best individual accuracy ($P_t - P_b$) for the 9 combination schemes are shown in Tables 1 to 4. We denote by P_{ia} the individual average of the team. The bar above P denotes the mean value over all generated teams for the respective experiment. Example histograms of $P_{AVR} - P_b$ and $P_{AVR} - P_w$ are given in Figure 1.

[1] Created by Dr. William H. Wolberg, W. Nick Street and Olvi L. Mangasarian, University of Wisconsin

[2] http://www.ics.uci.edu/ mlearn/MLRepository.html

Table 1: Results for the $\langle 4,4,2\rangle$ experiments, linear classifiers. $\bar{P}_w = 83.60$ %, $\bar{P}_{ia} = 87.62$ %, $\bar{P}_b = 90.43$ %

Combination scheme	min$(P_t - Pb)$ (in %)	max$(P_t - Pb)$ (in %)	% better than single best	% worse than single worst	% Correlation with $P_b - P_w$
MAJ	-3.51	2.81	32.13	0.13	3.46
NB	-3.51	3.16	42.03	0.06	8.42
BKS	-8.42	3.16	31.40	2.22	18.52
WER	-5.96	3.51	36.86	1.87	16.29
MAX	-4.56	3.51	37.21	0	-42.08
MIN	-4.56	3.51	37.21	0	-42.08
AVR	-4.21	3.51	46.67	0	-18.68
PRO	-4.21	2.81	44.38	0.13	-19.17
DT	-2.11	4.21	79.62	0	- 0.54

Table 2: Results for the $\langle 4,4,2\rangle$ experiments, quadratic classifiers. $\bar{P}_w = 85.35$ %, $\bar{P}_{ia} = 89.01$ %, $\bar{P}_b = 91.55$ %

Combination scheme	min$(P_t - Pb)$ (in %)	max$(P_t - Pb)$ (in %)	% better than single best	% worse than single worst	% Correlation with $P_b - P_w$
MAJ	-3.86	2.46	18.67	0.25	-6.53
NB	-3.86	2.46	18.79	0.25	-2.51
BKS	-4.21	2.46	21.56	1.59	6.39
WER	-4.21	3.16	19.21	1.24	7.18
MAX	-3.16	2.81	29.71	0.25	-3.37
MIN	-3.16	2.81	29.71	0.25	-3.37
AVR	-3.16	2.81	27.87	0.06	2.18
PRO	-3.16	2.46	28.00	0.25	0.20
DT	-2.81	2.46	35.75	0.25	4.98

Table 3: Results for the $\langle 4,3,3\rangle$ experiments, linear classifiers. $\bar{P}_w = 87.72$ %, $\bar{P}_{ia} = 90.46$ %, $\bar{P}_b = 92.42$ %.

Combination scheme	$\min(P_t - Pb)$ (in %)	$\max(P_t - Pb)$ (in %)	% better than single best	% worse than single worst	% Correlation with $P_b - P_w$
MAJ	-3.86	2.46	41.10	0	-17.16
NB	-3.86	2.46	39.29	0	-19.06
BKS	-7.72	2.46	20.26	3.26	-12.80
WER	-7.02	2.46	19.17	2.88	-17.53
MAX	-3.16	3.51	59.86	0	-58.55
MIN	-3.16	3.51	59.86	0	-58.55
AVR	-2.81	3.51	68.29	0	-33.07
PRO	-2.81	3.16	64.81	0	-41.23
DT	-1.75	4.21	86.67	0	-45.18

Table 4: Results for the $\langle 4,3,3\rangle$ experiments, quadratic classifiers. $\bar{P}_w = 88.43$ %, $\bar{P}_{ia} = 91.17$ %, $\bar{P}_b = 93.29$ %

Combination scheme	$\min(P_t - Pb)$ (in %)	$\max(P_t - Pb)$ (in %)	% better than single best	% worse than single worst	% Correlation with $P_b - P_w$
MAJ	-4.56	3.16	26.90	0.19	-18.31
NB	-4.56	3.16	26.90	0.19	-18.17
BKS	-5.96	3.51	24.24	3.55	-1.62
WER	-5.61	3.16	20.64	4.33	-0.52
MAX	-3.86	2.81	20.38	1.29	-9.21
MIN	-3.86	2.81	20.38	1.29	-9.21
AVR	-3.86	3.51	28.14	0.14	-2.51
PRO	-3.51	2.81	17.62	0.95	-1.16
DT	-3.86	3.16	23.48	0.43	-16.94

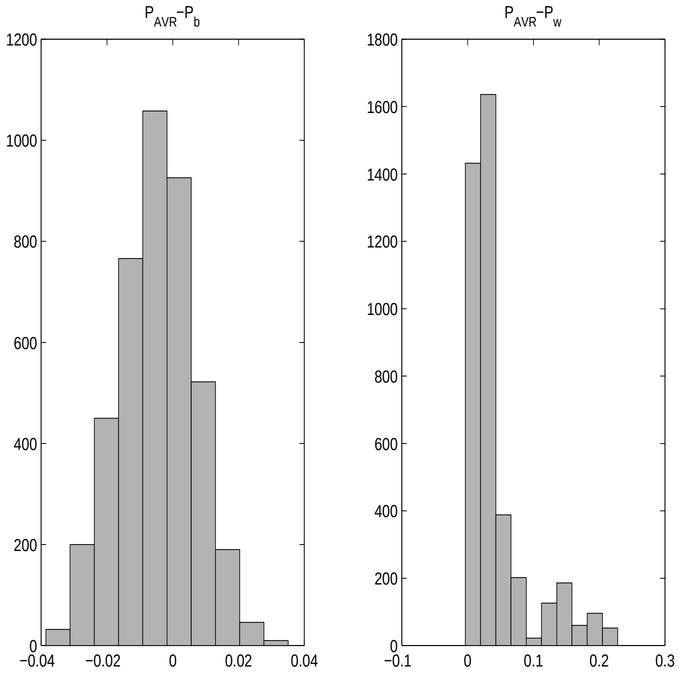


Figure 1: Histograms illustrating the distribution of the improvement for the Average aggregation method, experiment $\langle 4,3,3\rangle$ and quadratic individual classifiers.

2. Given in Tables 1 to 4 are the fraction of cases when $P_t > P_b$ (the probability that the team is better than the single best classifier) and also the fraction when $P_t < P_w$ (the team is worse than the worst classifier). As an illustration, Figure 2 plots the accuracy of the Decision Template combination, P_{DT} and the single best accuracy P_b versus the (sorted by P_{DT}) number of splits for experiment # 3.

3. The last columns in Tables 1 to 4 show the correlation between $P_b - P_w$ and P_t. Figure 3 displays an example of the relationship between P_{DT} and $P_b - P_w$.

4. A Two-way ANOVA was run to estimate whether there is a significant difference between the 9 combination schemes. The test found significant differences between the means of the team accuracies computed by the 9 schemes. The means with the 95 % confidence intervals from the $\langle 4,4,2\rangle$ experiments are shown in Figure 4 and from the $\langle 4,3,3\rangle$ experiments, in Figure 5.

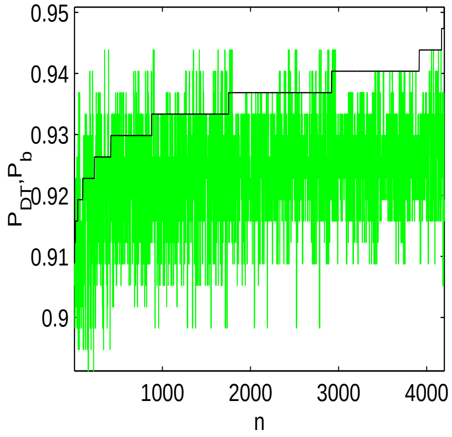


Figure 2: Sorted P_{DT} (the black line) and P_b (the grey line) for experiment $\langle 4, 3, 3\rangle$ and linear individual classifiers. $P_{DT} > P_b$ in 86.7 % of the cases.

4 Conclusions

1. *How is $P_t - P_b$ distributed and what are the maximal and the minimal possible values for different combination schemes?*

The difference between the team accuracy and the best individual shows a stable pattern. In all experiments the accuracy increases by a few per cent. The maximum of the $\max(P_t - Pb)$ in tables 1 to 4 is the Decision Template combination method with 4.21 % in Tables 1 and 3. All minimal values of $P_t - P_b$ are negative indicating that there is no combination scheme (at least not among the studied ones) that *guarantees* improvement over the single best classifier. The combination schemes with the worst negative result are the BKS and the Wernecke's method (up to -8.42 % for BKS). BKS is known for being prone to overtraining, so the result is not surprising. In general, the classification accuracy of the individual classifiers is around 90 %, so we cannot expect substantial improvement from the combination. The differences have approximately normal distributions for all combination schemes, a typical example is shown in the left plot in Figure 1. Results much worse than the single worst classifier are unlikely, as shown in the penultimate columns in Tables 1 to 4 which have very small (often zero) values. However, the distribution is not normal as is shown by the typical example in the right plot in Figure 1.

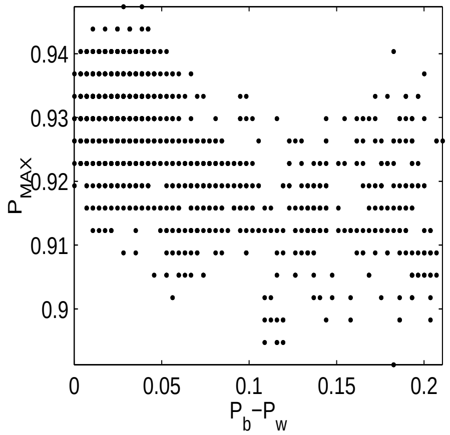
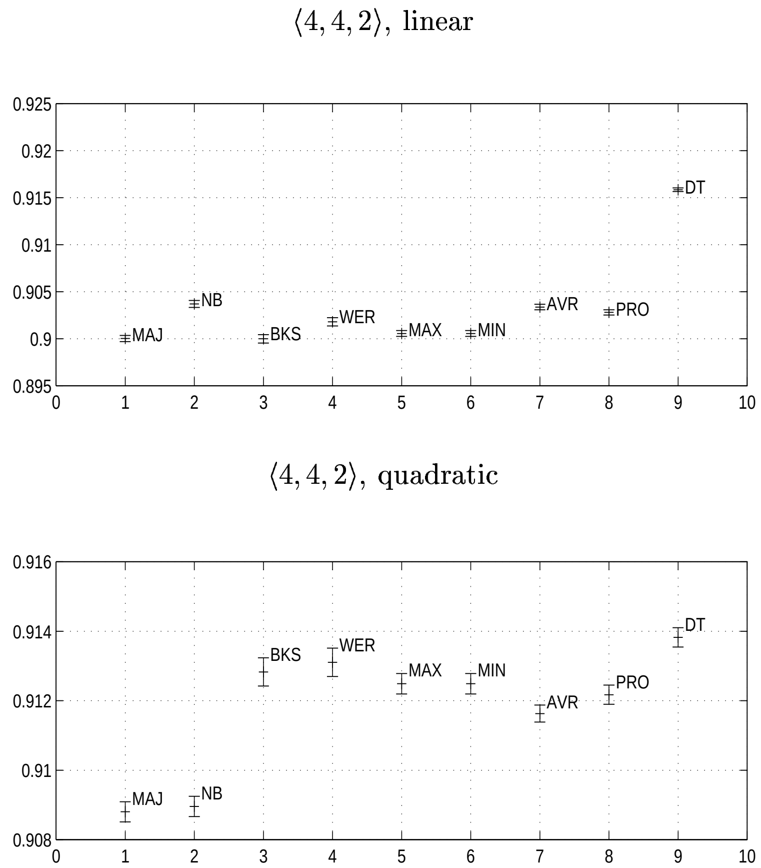


Figure 3: Scatterplot of P_{MAX} versus $P_b - P_w$ for $\langle 4,3,3 \rangle$ and linear individual classifiers. The correlation between the two is -0.59.

2. *How likely is an improvement* $(P_t - Pb > 0)$ *if we pick a random partition of the set of features?* The numerical answers to this question are given in the fourth column in Tables 1 to 4 for the experiments we carried out. However, we cannot offer a clear-cut conclusion. A persistent pattern is that the percentage getting an improvement over the single best dramatically depends on the quality of the individual classifiers. For the (weak) linear models, the improvement is more often encountered whereas for the quadratic models the chance for improvements are halved. For example, the DT combination has a chance of about 85 % (Table 3) to improve on the single best linear classifier if the feature set is split randomly into subsets of 4, 3, and 3 features. For the same split, when quadratic classifiers are used, none of the schemes has more than about 30 % chance for improvement (Table 4).

Perhaps we can conjecture that this chance depends on the problem (how complex it is), the classifier model (weak or strong), the number of features used per classifier, and more factors which we did not examine here, e.g., the number of classifiers L.

3. *Is the team accuracy* P_t *related to the range* $P_b - P_w$ *of individual accuracies?* The correlation coefficients, and the scatterplot in Figure 3 do not indicate an unequivocal relationship. As the correlation coefficients tend to be small by absolute value, there is little evidence that the more similar the individual accuracies the higher the improvement.



$\langle 4,4,2\rangle$, linear

$\langle 4,4,2\rangle$, quadratic

Figure 4: Means and the 95 % confidence intervals for experiment $\langle 4,4,2\rangle$.

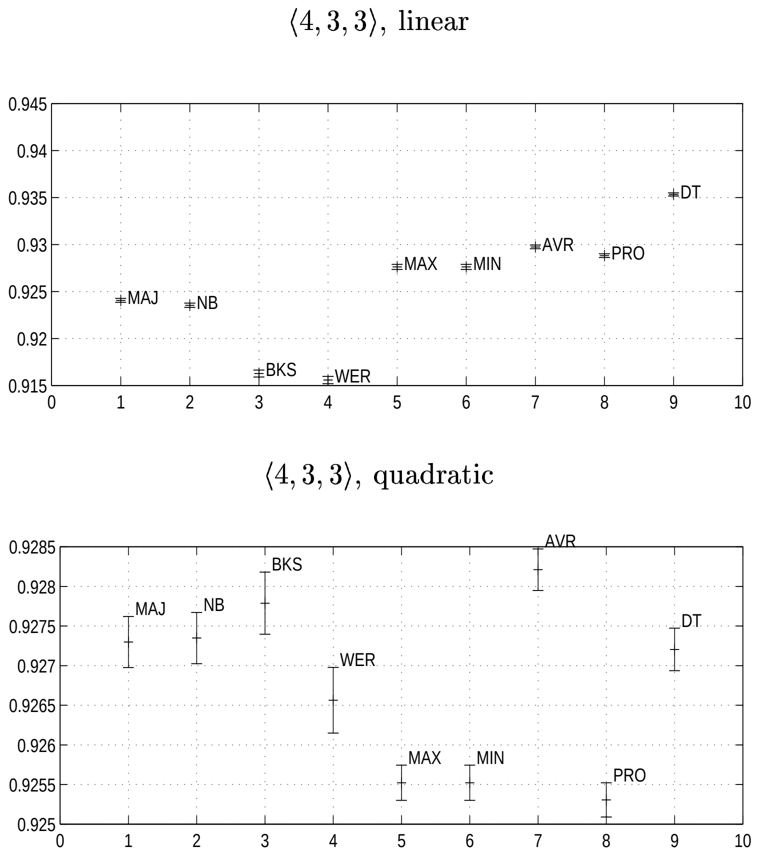
4. *How do the combination schemes compare with respect to the answers to the previous three questions?*

Again, the combination schemes exhibit variable performance, and this shows that: (a) there is no "best" combination for all scenarios, and (b) building a classifier team that outperforms the single best individual is a delicate job. Based on our results, we nominate the Decision Templates as the most successful combination scheme in our experiments.

References

[1] J.C. Bezdek, J.M Keller, R. Krishnapuram, and N.R. Pal. *Fuzzy Models and Algorithms for Pattern Recognition and Image Processing.* Kluwer Academic Publishers, 1999.

[2] K. Chen, L. Wang, and H. Chi. Methods of combining multiple classifiers with different features


⟨4, 3, 3⟩, linear

⟨4, 3, 3⟩, quadratic

Figure 5: Means and the 95 % confidence intervals for experiment $\langle 4, 3, 3\rangle$.

and their applications to text-independent speaker identification. *International Journal on Pattern Recognition and Artificial Intelligence*, 11(3):417–445, 1997.

[3] T.G. Dietterich. Ensemble methods in machine learning. In J. Kittler and F. Roli, editors, *Multiple Classifier Systems*, volume 1857 of *Lecture Notes in Computer Science*, pages 1–15, Cagliari, Italy, 2000. Springer.

[4] R.P.W. Duin and D.M.J. Tax. Experiments with classifier combination rules. In J. Kittler and F. Roli, editors, *Multiple Classifier Systems*, volume 1857 of *Lecture Notes in Computer Science*, pages 16–29, Cagliari, Italy, 2000. Springer.

[5] T.K. Ho. The random space method for constructing decision forests. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 20(8):832–844, 1998.

[6] Y.S. Huang and C.Y. Suen. A method of combining multiple experts for the recognition of unconstrained handwritten numerals. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 17:90–93, 1995.

[7] J. Kittler, M. Hatef, R.P.W. Duin, and J. Matas. On combining classifiers. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 20(3):226–239, 1998.

[8] J. Kittler, A. Hojjatoleslami, and T. Windeatt. Strategies for combining classifiers employing shared and distinct representations. *Pattern Recognition Letters*, 18:1373–1377, 1997.

[9] L.I. Kuncheva. Genetic algorithm for feature selection for parallel classifiers. *Information Processing Letters*, 46:163–168, 1993.

[10] L.I. Kuncheva, J.C. Bezdek, and R.P.W. Duin. Decision templates for multiple classifier fusion: an experimental comparison. *Pattern Recognition*, 1999. (accepted).

[11] L.I Kuncheva and L.C. Jain. Designing classifier fusion systems by genetic algorithms. *IEEE Transactions on Evolutionary Computation*, 4(4):327–336, 2000.

[12] L.I. Kuncheva, C.J. Whitaker, C.A. Shipp, and R.P.W. Duin. Limits on the majority vote accuracy in classifier fusion. (submitted).

[13] P. Latinne, O. Debeir, and C. Decaestecker. Different ways of weakening decision trees and their impact on classification accuracy of dt combination. In J. Kittler and F. Roli, editors, *Multiple Classifier Systems*, volume 1857 of *Lecture Notes in Computer Science*, pages 230–239, Cagliari, Italy, 2000. Springer.

[14] H.-S. Park and S.-W. Lee. Off-line recognition of large-set handwritten characters with multiple hidden Markov models. *Pattern Recognition*, 29(2):231–244, 1996.

[15] M. van Breukelen, R.P.W Duin, D.M.J. Tax, and J.E. den Hartog. Combining classifiers for the recognition of handwritten digits. In *I-st IAPR TC1 Workshop on Statistical Techniques in Pattern Recognition*, pages 13–18, Prague, Czech Republic, 1997.

[16] K.-D. Wernecke. A coupling procedure for discrimination of mixed data. *Biometrics*, 48:497–506, 1992.

[17] L. Xu, A. Krzyzak, and C.Y. Suen. Methods of combining multiple classifiers and their application to handwriting recognition. *IEEE Transactions on Systems, Man, and Cybernetics*, 22:418–435, 1992.