A Theoretical Study on Six Classifier Fusion Strategies

Ludmila I. Kuncheva, *Member*, *IEEE*

Abstract—We look at a single point in the feature space, two classes, and L classifiers estimating the posterior probability for class ω_1. Assuming that the estimates are independent and identically distributed (normal or uniform), we give formulas for the classification error for the following fusion methods: average, minimum, maximum, median, majority vote, and oracle.

Index Terms—Classifier combination, theoretical error, fusion methods, order statistics, majority vote, independent classifiers.

1 Introduction

Classifier combination has received considerable attention in the past decade and is now an established pattern recognition offspring. Recently, the focus has been shifting from practical heuristic solutions of the combination problem toward explaining why combination methods and strategies work so well and in what cases some methods are better than others.

Let $\mathcal{D} = \{D_1, \ldots, D_L\}$ be a set (pool/committee/ensemble/team) of classifiers. By combining the individual outputs, we aim at a higher accuracy than that of the best classifier. There is a consensus among the researchers in classifier combination that the major factor for a better accuracy is the diversity in the classifier team and, so, the fusion method is of a secondary importance (see [10]). However, a choice of an appropriate fusion method can improve further on the performance of the ensemble.

This study was inspired by a recent publication by Alkoot and Kittler [1] where classifier fusion methods are experimentally compared. Here, we look at a theoretical backup of some of the results. Let $\Re^n$ be the feature space. As in [1], we make the following **assumptions**:

1. All classifiers produce soft class labels. We assume that $d_{j,i}(\mathbf{x}) \in [0,1]$ is an estimate of the posterior probability $P(\omega_i|\mathbf{x})$ offered by classifier D_j for an input $\mathbf{x} \in \Re^n$, $i = 1, 2,\ \ j = 1, \ldots, L$.
2. There are two possible classes $\Omega = \{\omega_1, \omega_2\}$. We consider the case where, for any $\mathbf{x}$, $d_{j,1}(\mathbf{x}) + d_{j,2}(\mathbf{x}) = 1$, $j = 1, \ldots, L$.
3. A single point $\mathbf{x} \in \Re^n$ is considered and the true posterior probability is $P(\omega_1|\mathbf{x}) = p > 0.5$. Thus, the Bayes-optimal class label for $\mathbf{x}$ is ω_1 and a classification error occurs if label ω_2 is assigned.
4. The classifiers commit independent and identically distributed errors in estimating $P(\omega_1|\mathbf{x})$.

Two distributions of $d_{j,1}(\mathbf{x})$ have been discussed in [1]: a normal distribution with mean p and variance σ^2 (σ varied between 0.1 and 1) and a uniform distribution spanning the interval $[p-b, p+b]$ (b varied from 0.1 to 1).

Simple fusion methods are the most obvious choice when constructing a multiple classifier system [4], [5], [11], [13], [14], i.e., the support for class ω_i, $d_i(\mathbf{x})$, yielded by the team is

$$d_i(\mathbf{x}) = \mathcal{F}(d_{1,i}(\mathbf{x}), \ldots, d_{L,i}(\mathbf{x})), \quad i = 1, 2, \tag{1}$$

where $\mathcal{F}$ is the chosen fusion method. Here, we study the fusion methods compared in [1], except the product, i.e., $\mathcal{F}$ stands for

- minimum
- maximum
- average
- median
- majority vote

For the majority vote, we first "harden" the individual decisions by assigning class labels $D_j(\mathbf{x}) = \omega_1$ if $d_{j,1}(\mathbf{x}) > 0.5$, and $D_j(\mathbf{x}) = \omega_2$ if $d_{j,1}(\mathbf{x}) \leq 0.5$, $j = 1, \ldots, L$. Then, the class label most represented among the L (label) outputs is chosen.

The reason to leave the product out was that it did not fit easily within the theoretical framework proposed here. We have added an abstract fusion model, called the *oracle*. In this model, if at least one of the classifiers produces the correct class label, then the team produces the correct class label too. Oracle is usually used in comparative experiments.

The rest of the paper is organized as follows: Section 2 formalizes the relationship between the probability of error of the team, the parameters of the distributions (b and σ), the true posterior probability p, and the number of classifiers L. Section 3 reproduces part of the experiments from [1] and Section 4 concludes the study.

2 Probability of Error for the Selected Fusion Methods

2.1 The Two Distributions

Denote by P_j the output of classifier D_j for class ω_1, i.e., $P_j = d_{j,1}(\mathbf{x})$ and let

$$\hat{P}_1 = \mathcal{F}(P_1, \ldots, P_L) \tag{2}$$

be the fused estimate of $P(\omega_1|\mathbf{x})$. By assumption 2, the posterior probability estimates for ω_2 are $1 - P_j$, $j = 1, \ldots, L$. The same fusion method $\mathcal{F}$ is used to find the fused estimate of $P(\omega_2|\mathbf{x})$,

$$\hat{P}_2 = \mathcal{F}(1 - P_1, \ldots, 1 - P_L). \tag{3}$$

We regard the individual estimates P_j as independent identically distributed random variables, such that $P_j = p + \epsilon_j$, with probability density functions (pdf) $f(y), y \in \Re$ and cumulative distribution functions (cdf) $F(t), t \in \Re$. Then, $\hat{P}_1$ is a random variable too with a pdf $f_{\hat{P}_1}(y)$ and cdf $F_{\hat{P}_1}(t)$.

The single classifier, the average, and the median fusion models will result in $\hat{P}_1 + \hat{P}_2 = 1$. The higher of the two estimates determines the class label. The oracle and the majority vote make decisions on the class label outputs and we can stipulate that $\hat{P}_1 = 1, \hat{P}_2 = 0$ if class ω_1 is assigned and $\hat{P}_1 = 0, \hat{P}_2 = 1$ for class ω_2. Thus, it is necessary and sufficient to have $\hat{P}_1 > 0.5$ to label $\mathbf{x}$ in ω_1 (the correct label). The probability of error, given $\mathbf{x}$, denoted P_e, is

$$P_e = P(\text{error}|\mathbf{x}) = P(\hat{P}_1 \leq 0.5) = F_{\hat{P}_1}(0.5) = \int_0^{0.5} f_{\hat{P}_1}(y)dy \tag{4}$$

for the single best classifier, average, median, majority vote, and the oracle.

For the minimum and the maximum rules, however, the sum of the fused estimates is not necessarily one. The class label is then decided by the maximum of $\hat{P}_1$ and $\hat{P}_2$. Thus, an error will occur if $\hat{P}_1 \leq \hat{P}_2$,[1]

• *The author is with the School of Informatics, University of Wales, Bangor Dean Street, Bangor, Gwynedd LL57 1UT, UK. E-mail: l.i.kuncheva@bangor.ac.uk.*

Manuscript received 10 July 2000; revised 21 May 2001; accepted 27 July 2001.
Recommended for acceptance by T.K. Ho.
For information on obtaining reprints of this article, please send e-mail to: tpami@computer.org, and reference IEEECS Log Number 112423.

1. We note that since P_1 and P_2 are continuous-valued random variables, the inequalities can be written with or without the equal sign, i.e., $\hat{P}_1 > 0.5$ is equivalent to $\hat{P}_1 \geq 0.5$, etc.

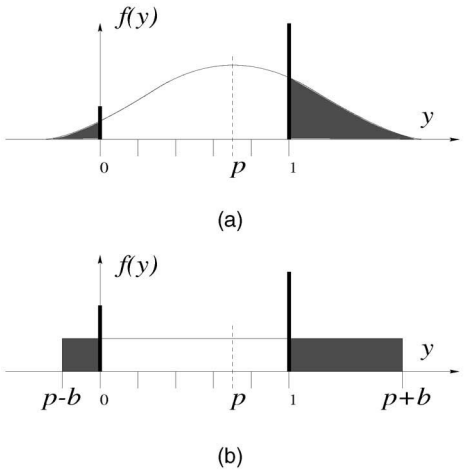

Fig. 1. "Clipped" distributions: (a) normal, $\sim N(p, \sigma^2)$; (b) uniform within $[p-b, p+b]$. The probabilities corresponding to the shaded areas are assigned to the boundry values 0 and 1.

$$P_e = P(\text{error}|\mathbf{x}) = P(\hat{P}_1 \leq \hat{P}_2) \tag{5}$$

for the minimum and the maximum.

Two distributions of the P_j's are considered:

- Normal distribution, $N(p, \sigma^2)$. We denote by $\Phi(z)$ the cumulative distribution function of $N(0,1)$.[2] Thus, the cumulative distribution function for the normal distribution considered here is

$$F(t) = \Phi\left(\frac{t-p}{\sigma}\right). \tag{6}$$

- Uniform distribution within $[p-b, p+b]$, i.e.,

$$f(y) = \begin{cases} \frac{1}{2b}, & y \in [p-b, p+b]; \\ 0, & \text{elsewhere}, \end{cases}$$
$$F(t) = \begin{cases} 0, & t \in (-\infty, p-b); \\ \frac{t-p+b}{2b}, & t \in [p-b, p+b]; \\ 1, & t > p+b. \end{cases} \tag{7}$$

In [1], the distributions are "clipped" so that all P_js were in $[0,1]$, as illustrated in Fig. 1. In the simulation process from the original distributions, negative values of P_j were replaced by 0s and values above 1 were replaced by by 1s. Thus, the true distributions examined in [1] are as tabulated in Tables 1 and 2.

A theoretical analysis with the clipped distributions is not straightforward. The clipped distributions are actually mixtures of a continuous random variable in the interval (0,1) and a discrete one taking values 0 or 1. This results in the "jumps" of the respective cumulative functions $F(t)$, as shown in Tables 1 and 2.

We can offer the following argument for using the original distributions. Suppose that p is not a probability but *the amount of support* for ω_1. The support for ω_2 will be again $1-p$. In estimating p, we do not have to restrict P_js within the interval $[0,1]$. For example, a neural network (or *any* classifier for that matter) trained by minimizing the squared error between its output and the zero-one (class label) target function produces an estimate of the posterior probability for that class (cf. [2]). Thus, depending on the parameters and the transition functions, a neural network output (that approximates p) might be greater than 1 or even negative. We take the L values (in $\Re$) and fuse them by (2) and (3) to get $\hat{P}_1$. The same rule applies, i.e., ω_1 is assigned by the team if $\hat{P}_1 > \hat{P}_2$. Then,

2. Available in tabulated form or from any statistical package.

TABLE 1
The Clipped-Normal Distribution Function $F(t)$

Interval for t	$(-\infty, 0)$	$[0, 1)$	$[1, \infty)$
$F(t)$	0	$\Phi\left(\frac{t-p}{\sigma}\right)$	1

we calculate the *probability* of error P_e as $P(\hat{P}_1 \leq \hat{P}_2)$. This calculation does not require in any way that P_js are probabilities or are within the unit interval. Therefore, the original (nonclipped) distributions of the estimates P_j can be used.

Clipping the distributions may have affected the calculation of the empirical error rate in [1]. For example, for large bs, many of the P_js will be either 0s or 1s. If there are indices j_1 and j_2 such that $P_{j_1} = 1$ and $P_{j_2} = 0$, then both the minimum and the maximum will produce a tie, and the random tie break will induce bias in P_e. This can explain some discrepancy with the results from our study (relegated to Section 4). For small b and σ, however, where P_js are mostly or entirely in $[0,1]$, we shall expect compatible results.

2.2 Single Classifier

Since $F_{\hat{P}_1}(t) = F(t)$, the error of a single classifier for the normal distribution is

$$P_e = \Phi\left(\frac{0.5-p}{\sigma}\right), \tag{8}$$

and for the uniform distribution,

$$P_e = \frac{0.5-p+b}{2b}. \tag{9}$$

2.3 Minimum and Maximum

These two fusion methods are considered together because, as shown below, they are identical for $c = 2$ classes and any number of classifiers L, where P_j and $1-P_j$ are the estimates by classifier D_j for $P(\omega_1|\mathbf{x})$ and $P(\omega_2|\mathbf{x})$, respectively.

Substituting $\mathcal{F} = \max$ in (2), the team's support for ω_1 is $\hat{P}_1 = \max_j\{P_j\}$. Therefore, the support for ω_2 is $\hat{P}_2 = \max_j\{1-P_j\}$. A classification error will occur if

$$\max_j\{P_j\} < \max_j\{1-P_j\}, \tag{10}$$
$$p + \max_j\{\epsilon_j\} < 1 - p - \min_j\{\epsilon_j\}, \tag{11}$$
$$\epsilon_{\max} + \epsilon_{\min} < 1 - 2p. \tag{12}$$

For the minimum fusion method, an error will occur if

$$\min_j\{P_j\} < \min_j\{1-P_j\}, \tag{13}$$
$$p + \epsilon_{\min} < 1 - p - \epsilon_{\max}, \tag{14}$$
$$\epsilon_{\max} + \epsilon_{\min} < 1 - 2p, \tag{15}$$

which proves the equivalence. Notice that the assumption of independence has not been used for this equivalence, nor have the types of the distributions of $\hat{P}_j$s.

TABLE 2
The Clipped-Uniform Distribution Function $F(t)$

Interval for t	$(-\infty, t_1)$	$[t_1, t_2)$	$[t_2, \infty)$
$F(t)$	0	$\frac{t-p+b}{2b}$	1

Notations: $t_1 = \max\{0, p-b\}$ *and* $t_2 = \min\{1, p+b\}$.

TABLE 3
The Theoretical Error P_e for the Single Classifier and the Six Fusion Methods

Method	P_e for Normal distribution	P_e for Uniform distribution ($p-b<0.5$)
Single classifier	$\Phi\left(\frac{0.5-p}{\sigma}\right)$	$\frac{0.5-p+b}{2b}$
Minimum/Maximum		$\frac{1}{2}\left(\frac{1-2p}{2b}+1\right)^L$
Average	$\Phi\left(\frac{\sqrt{L}(0.5-p)}{\sigma}\right)$	$\Phi\left(\frac{\sqrt{3L}(0.5-p)}{b}\right)$
Median/Vote	$\sum_{j=\frac{L+1}{2}}^{L}\binom{L}{j}\Phi\left(\frac{0.5-p}{\sigma}\right)^j\left[1-\Phi\left(\frac{0.5-p}{\sigma}\right)\right]^{L-j}$	$\sum_{j=\frac{L+1}{2}}^{L}\binom{L}{j}\left(\frac{0.5-p+b}{2b}\right)^j\left[1-\frac{0.5-p+b}{2b}\right]^{L-j}$
Oracle	$\Phi\left(\frac{0.5-p}{\sigma}\right)^L$	$\left(\frac{0.5-p+b}{2b}\right)^L$

The probability of error for minimum and maximum is

$$P_e = P(\epsilon_{\max}+\epsilon_{\min}<1-2p) \quad (16)$$
$$= F_{\epsilon_s}(1-2p), \quad (17)$$

where $F_{\epsilon_s}(t)$ is the cdf of the random variable $s=\epsilon_{\max}+\epsilon_{\min}$. For the normally distributed P_js, ϵ_j are also normally distributed with mean 0 and variance σ^2. However, we cannot assume that $\epsilon_{\max}$ and $\epsilon_{\min}$ are independent and analyze their sum as another normally distributed variable because these are *order statistics* and $\epsilon_{\min}\leq\epsilon_{\max}$. We have not attempted a solution for the normal distribution case.

For the uniform distribution, we follow an example taken from [8], where the pdf of the midrange $(\epsilon_{\min}+\epsilon_{\max})/2$ is calculated for L observations. We derived $F_{\epsilon_s}(t)$ to be

$$F_{\epsilon_s}(t)=\begin{cases}\frac{1}{2}\left(\frac{t}{2b}+1\right)^L, & t\in[-2b,0];\\ 1-\frac{1}{2}\left(1-\frac{t}{2b}\right)^L, & t\in[0,2b].\end{cases} \quad (18)$$

Noting that $t=1-2p$ is always negative,

$$P_e = F_{\epsilon_s}(1-2p) = \frac{1}{2}\left(\frac{1-2p}{2b}+1\right)^L. \quad (19)$$

2.4 Average

The average fusion method gives $\hat{P}_1=\frac{1}{L}\sum_{j=1}^{L}P_j$. If $P_1,\ldots,P_L$ are normally distributed (and independent!), then $\hat{P}\sim N\left(p,\frac{\sigma^2}{L}\right)$. The probability of error for this case is

$$P_e = P(\hat{P}_1<0.5)=\Phi\left(\frac{\sqrt{L}(0.5-p)}{\sigma}\right). \quad (20)$$

The calculation of P_e for the case of uniform distribution is not that straightforward. We can assume that the sum of L independent variables will result in a variable of approximately normal distribution. The higher the L, the more accurate the approximation. Knowing that the variance of the uniform distribution for P_j is $\frac{b^2}{3}$, we can assume $\hat{P}\sim N\left(p,\frac{b^2}{3L}\right)$. Then,

$$P_e = P(\hat{P}_1<0.5)=\Phi\left(\frac{\sqrt{3L}(0.5-p)}{b}\right). \quad (21)$$

2.5 Median and Majority Vote

These two fusion methods are pooled because they are identical for the current setup.

Since only two classes are considered, we restrict our choice of L to odd numbers only. An even L is inconvenient for at least two reasons. First, the majority vote might tie. Second, the theoretical analysis of a median which is calculated as the average of the $(L/2)$ and $(L/2+1)$ order statistics is cumbersome.

For the median fusion method,

$$\hat{P}_1=\text{med}\{P_1,\ldots,P_L\}=p+\text{med}\{\epsilon_1,\ldots,\epsilon_L\}=p+\epsilon_m. \quad (22)$$

Then, the probability of error is

$$P_e = P(p+\epsilon_m<0.5)=P(\epsilon_m<0.5-p)=F_{\epsilon_m}(0.5-p), \quad (23)$$

where F_{ϵ_m} is the cdf of ϵ_m. From the order statistics theory [8],

$$F_{\epsilon_m}(t)=\sum_{j=\frac{L+1}{2}}^{L}\binom{L}{j}F_\epsilon(t)^j[1-F_\epsilon(t)]^{L-j}, \quad (24)$$

where $F_\epsilon(t)$ is the distribution of ϵ_j, i.e., $N(0,\sigma^2)$ or uniform in $[-b,b]$. We can now substitute the two cdf, to obtain the respective P_e

- for the normal distribution

$$P_e=\sum_{j=\frac{L+1}{2}}^{L}\binom{L}{j}\Phi\left(\frac{0.5-p}{\sigma}\right)^j\left[1-\Phi\left(\frac{0.5-p}{\sigma}\right)\right]^{L-j}. \quad (25)$$

- for the uniform distribution

$$P_e=\begin{cases}0, & p-b>0.5;\\ \sum_{j=\frac{L+1}{2}}^{L}\binom{L}{j}\left(\frac{0.5-p+b}{2b}\right)^j\left[1-\frac{0.5-p+b}{2b}\right]^{L-j}, & \text{otherwise.}\end{cases} \quad (26)$$

The majority vote will assign the wrong class label, ω_2, to $\mathbf{x}$ if at least $\frac{L+1}{2}$ classifiers vote for ω_2. The probability that a single classifier is wrong is given by (8) for the normal distribution and (9) for the uniform distribution. Denote this probability by P_s. Since the classifiers are independent, the probability that at least $\frac{L+1}{2}$ are wrong is calculated by the binomial formula

$$P_e=\sum_{j=\frac{L+1}{2}}^{L}\binom{L}{j}P_s^j[1-P_s]^{L-j}. \quad (27)$$

By substituting for P_s from (8) and (9), we recover (25) and (26) for the normal and the uniform distribution, respectively.

2.6 The Oracle

The probability of error for the oracle is

$$P_e = P(\text{all incorrect})=F(0.5)^L \quad (28)$$

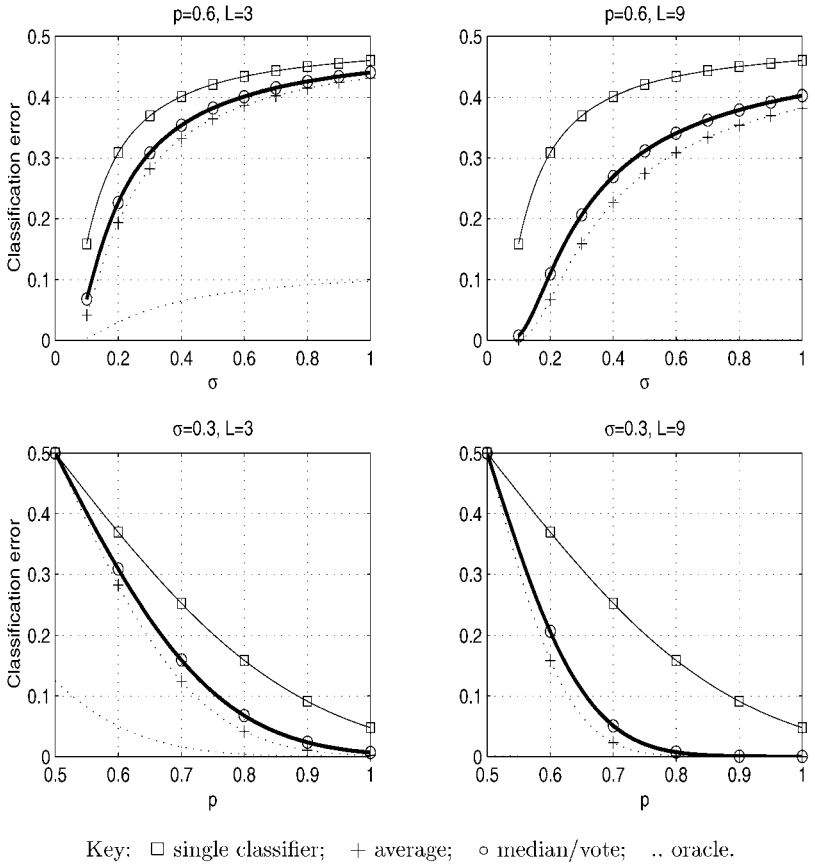

Fig. 2. P_e for **normally** distributed P_js.

For the normal distribution

$$P_e = \Phi\left(\frac{0.5-p}{\sigma}\right)^L, \tag{29}$$

and for the uniform distribution

$$P_e = \begin{cases} 0, & p-b>0.5; \\ \left(\frac{0.5-p+b}{2b}\right)^L, & \text{otherwise.} \end{cases} \tag{30}$$

Table 3 displays in a compact form the results for the two distributions, the single classifier, and the six fusion methods.

3 Illustration Example

A direct comparison between the errors in Table 3 is hardly possible, except for the single classifier and the oracle, where the preference is known anyway. In this section, we reproduce part of the experiments from [1]. Fig. 2 plots the classification error of the single classifier and the team, calculated by the respective equations for normally distributed P_js. The top two plots depict P_e against σ for a fixed $p = 0.6$ and the bottom two plots depict P_e against p for a fixed $\sigma = 0.3$. Fig. 3 displays the results for uniformly distributed P_js. The top two plots depict P_e against b for a fixed $p = 0.6$ and the bottom two plots depict P_e against p for a fixed $b = 0.8$. (The fixed values are the same as in [1].)

The results can be summarized as

1. Expected results. These are well-documented in the literature on classifier combination.
   a. The individual error is higher than the error of any of the fusion methods.
   b. The oracle model (an abstraction) is the best of all. For $L = 9$, the oracle error rate is approximately zero.
   c. The more classifiers we have in the team, the lower the error. Recall that the classifiers are assumed to be independent, which can hardly be achieved in real problems.
2. More interesting findings from this example.
   a. The average and the median/vote methods have approximately the same performance for normally distributed P_j but are different for the uniform distribution, the average being the better of the two.
   b. Contrary to some experimental evidence published elsewhere, the average method is outperformed by the minimum/maximum method. This observation is based on the uniform distribution model only. Unfortunately, the theoretical calculation of P_e for the minimum/maximum method in the case of normally distributed P_js is not easy, and we cannot draw a parallel with the average in this case.

Alkoot and Kittler's experimental results [1] differ from our results. For example, they found a threshold value for b, where minimum, maximum, and product change from the best to the worst fusion methods, even worse than the single classifier. Such change of behavior was not found here and the discrepancy can be attributed to the clipped-distributions effect. Also, we used

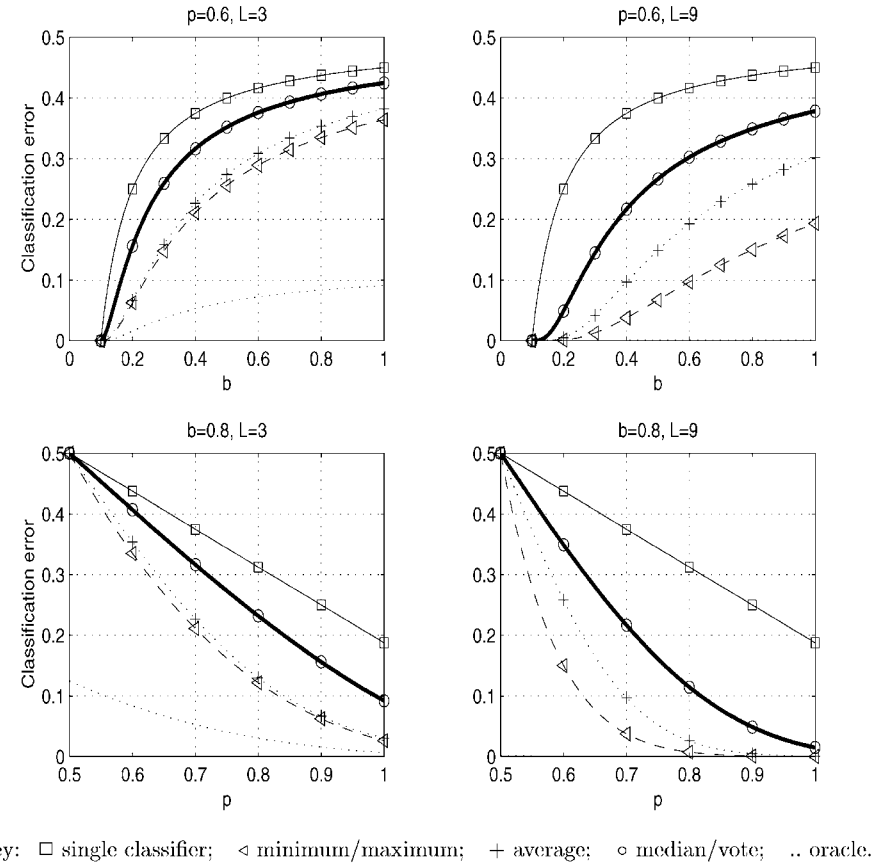

Fig. 3. P_e for **uniformly** distributed P_js.

$L = 9$ classifiers instead of $L = 8$ to avoid ties in the majority vote method, which explains the differences in the results. It should be noted that for small b and σ, i.e., when the distributions did not need to be clipped or the probability outside $[0, 1]$ was small, the two sets of results are similar.

4 Conclusions

Six simple classifier fusion methods have been studied theoretically: minimum, maximum, average, median, majority vote, and the oracle, together with the single classifier. We give formulas for the classification error at a single point in the feature space $\mathbf{x} \in \Re^n$ under the following conditions: two classes $\{\omega_1, \omega_2\}$; each classifier gives an output P_j as an estimate of the posterior probability $P(\omega_1|\mathbf{x}) = p > 0.5$; P_j are i.i.d coming from a fixed distribution (normal or uniform) with mean p. The formulas for P_e were derived for all cases, except for minimum and maximum fusion for normal distribution. It was shown that minimum and maximum are identical for $c = 2$ classes, regardless of the distribution of P_js and so are the median and the majority vote for the current set up. An illustration example was given, reproducing part of the experiments in [1]. Minimum/maximum fusion was found to be the best for uniformly distributed P_js (and did not participate in the comparison for the normal distribution).

It can be noted that for c classes, it is not enough that $P(\omega_1|\mathbf{x}) > \frac{1}{c}$ for a correct classification. For example, let $c = 3$ and let $P(\omega_1|\mathbf{x}) = 0.35, P(\omega_2|\mathbf{x}) = 0.45$, and $P(\omega_3|\mathbf{x}) = 0.20$. Then, although $P(\omega_1|\mathbf{x}) > \frac{1}{c}$, the class label will be ω_2. Therefore, only $P_1, \ldots, P_L$ are not enough and we also need to specify conditions for the support for the other classes. Conversely, $P(\omega_1|\mathbf{x}) > 0.5$ is sufficient but not necessary for a correct classification, and the true classification error can only be smaller than P_e. We have confined the comparative study to $c = 2$ because the analysis of the relative performance of the fusion methods when $c > 2$ requires making hypotheses about more variables. Our results differ at some places from the empirical findings in [1]. For example, we found that the minimum/maximum method is consistently better than the other fusion methods for the uniform distribution, whereas Alkoot and Kittler find a change of ranking for a certain b. This discrepancy in the results can be explained by the clipped-distribution model adopted in [1].

It is claimed in the literature that combination methods are less important than the diversity of the team. However, given a set of classifiers, the only way to extract the most of it is to pick a good combination method. Indeed, for normally distributed errors, the fusion methods gave very similar performance, but, for the uniformly distributed error, the methods differed significantly, especially for higher L. For example, the top right plot in Fig. 3 shows that, for $b = 0.5$, P_e for a single classifier of 40.0 percent can be reduced to 26.7 percent by the median or majority vote, 14.9 percent by the average, and 6.7 percent my the minimum or maximum fusion. This example comes in support of the idea that combination methods are also relevant in combining classifiers.

Similar studies can be carried out for distributions other than normal or uniform. Typically, we have little knowledge of the behavior of the classifier outputs and, so, the normal distribution is a natural choice.

At this stage, theoretical results (including this study) rely on numerous assumptions [4], [12]. Building a general theoretical framework for classifier fusion is a fascinating perspective, but it has to start somewhere. The lack of a more general theory is an indirect indication of the difficulties encountered. Indeed, even simple models lead to elaborate derivations under restrictive assumptions. This correspondence should be considered as a step toward this more general framework.

The most restrictive and admittedly unrealistic assumption is the independence of the estimates. It is recognized that "independently built" classifiers exhibit positive correlation [7], and this is attributed to the fact that difficult parts of the feature space are difficult for all classifiers. Ensemble design methods such as ADAboost or arcing [3], [9] try to overcome this unfavorable dependency by enforcing diversity. However, it is difficult to measure or express this diversity in a mathematically tractable way [6]. While statistical independence is rigorously defined, how are patterns of dependency expressed? Conceptualizing and quantifying diversity between classifier outputs is a challenging task on its own and will add a whole new dimension to classifier fusion.

REFERENCES

[1] F. Alkoot and J. Kittler, "Experimental Evaluation of Expert Fusion Strategies," *Pattern Recognition Letters*, vol. 20, pp. 1361–1369, 1999.

[2] C. Bishop, *Neural Networks for Pattern Recognition*. Oxford, UK: Clarendon Press, 1995.

[3] L. Breiman, "Combining Predictors," *Combining Artificial Neural Nets*, A. Sharkey, ed., pp. 31–50, 1999.

[4] J. Kittler, M. Hatef, R. Duin, and J. Matas, "On Combining Classifiers," *IEEE Trans. Pattern Analysis and Machine Intelligence*, vol. 20, no. 3, pp. 226–239, 1998.

[5] L. Kuncheva, J. Bezdek, and R. Duin, "Decision Templates for Multiple Classifier Fusion: An Experimental Comparison," *Pattern Recognition*, vol. 34, no. 2, pp. 299–314, 2001.

[6] L. Kuncheva and C. Whitaker, "Ten Measures of Diversity in Classifier Ensembles: Limits for Two Classifiers," *Proc. IEE Workshop Intelligent Sensor Processing*, pp. 10/1–10/6, Feb. 2001.

[7] B. Littlewood and D. Miller, "Conceptual Modeling of Coincident Failures in Multiversion Software," *IEEE Trans. Software Eng.*, vol. 15 no. 12, pp. 1596–1614, Dec. 1989.

[8] A. Mood, F. Graybill, and D. Boes, *Introduction to the Theory of Statistics*, third ed. McGraw-Hill, 1974.

[9] R. Schapire, "Theoretical Views of Boosting," *Proc. Fourth European Conf. Computational Learning Theory*, pp. 1–10, 1999.

[10] *Combining Artificial Neural Nets. Ensemble and Modular Multi-Net Systems*. A. Sharkey, ed., London: Springer-Verlag, 1999.

[11] D. Tax, R. Duin, and M. van Breukelen, "Comparison between Product and Mean Classifier Combination Rules," *Proc. Workshop Statistical Pattern Recognition*, 1997.

[12] K. Tumer and J. Ghosh, "Error Correlation and Error Reduction in Ensemble Classifiers," *Connection Science*, vol. 8, nos. 3 and 4, pp. 385–404, 1996.

[13] K. Tumer and J. Ghosh, "Linear and Order Statistics Combiners for Pattern Classification," *Combining Artificial Neural Nets*, A. Sharkey, ed., pp. 127–161, 1999.

[14] M. van Breukelen, R. Duin, D. Tax, and J. den Hartog, "Combining Classifiers for the Recognition of Handwritten Digits," *Proc. First IAPR TC1 Workshop Statistical Techniques in Pattern Recognition*, pp. 13–18, 1997.