Short Papers

Sum versus Vote Fusion in Multiple Classifier Systems

J. Kittler and F.M. Alkoot

Abstract—Amidst the conflicting experimental evidence of superiority of one over the other, we investigate the Sum and majority Vote combining rules in a two class case, under the assumption of experts being of equal strength and estimation errors conditionally independent and identically distributed. We show, analytically, that, for Gaussian estimation error distributions, Sum always outperforms Vote. For heavy tail distributions, we demonstrate by simulation that Vote may outperform Sum. Results on synthetic data confirm the theoretical predictions. Experiments on real data support the general findings, but also show the effect of the usual assumptions of conditional independence, identical error distributions, and common target outputs of the experts not being fully satisfied.

Index Terms—Multiple classifiers, fusion rules, estimation error.

1 Introduction

AMONG the many multiple classifier combination rules suggested in the literature [2], [4], [5], [6], [7], [8], [9], [11], [12], [15], Sum and Vote are used the most frequently. The Sum rule operates directly on the soft outputs of individual experts for each class hypothesis, normally delivered in terms of a posteriori class probabilities. The fused decision is obtained by applying the maximum value selector to the class dependent averages of these outputs. Vote, on the other hand, operates on class labels assigned to each pattern by the respective experts. The labels are obtained by hardening the soft decision outputs using the maximum value selector. The Vote rule output is a function of the votes received for each class from each single expert.

Classification systems implementing the Bayes decision rule incur classification errors over and above the Bayes rate due to errors in their estimates of the aposteriori class probabilities. The larger the variance of the error distribution, the larger the additional classification error. A multiple classifier system which deploys the Sum rule reduces this variance and, as a result, diminishes the additional classification error. The properties of the rule have been widely investigated [6], [7], [8], [13], [14]. As for the majority vote (Vote), Lam and Suen [9] give a comprehensive analysis of the rule under the assumption of conditional independence of the experts.

From the extensive studies of Sum and Vote reported in the literature, one would expect the former to outperform the latter because of the use of soft expert outputs. However, Vote has the important advantage that it can be applied directly even when individual experts do not output a posteriori class probabilities. Yet the experimental evidence in support of the relative performance of the two rules is conflicting. For instance, in [1], [8], it was found that Sum outperforms Vote, while, in [5], Vote performed better than Sum. The aim of this paper is to investigate the relationship between these two rules in more detail, under the assumption of the experts being of equal strength and estimation errors conditionally independent and identically distributed. As a main contribution of this paper, we demonstrate that the relative merits of Sum and Vote depend on the distribution of estimation errors. We show, analytically, in Section 2, that, for normally distributed estimation errors, Sum always outperforms Vote. But, for heavy tail distributions, that is, for error distributions with a significant probability mass in their tails, Vote may outperform Sum. Typically, such distributions are produced by classifiers based on Hidden Markov Models (HMM) [8]. We demonstrate this by experiments on synthetic data in Section 3. Interestingly, for heavy tail distributions, the superiority of Sum may be eroded for any number of experts if the margin between the two a posteriori class probabilities is small or for a small number of cooperating experts, even when the margin is large. In the latter case, once the number of experts exceeds a certain threshold, Sum tends to be superior to Vote. Experiments on real data presented in Section 4, which involve an HMM expert, support the general findings, but also show the effect of the assumptions of conditional independence, identical error distributions, and common target outputs of the experts not being fully satisfied.

The results presented in this paper, first of all, contribute to the understanding of these two fusion rules. Their practical relevance is particularly important when one or more experts fused is known to exhibit heavy tail estimation errors as this knowledge may favor the choice of Vote over the Sum rule.

2 Theoretical Analysis

Consider a two class pattern recognition problem. Let us assume that we have N classifiers, each representing a given pattern by an identical measurement vector $\mathbf{x}$. Now, according to the Bayesian decision theory, a pattern $\mathbf{x}$ should be assigned to class ω_j for which the a posteriori probability $P(\omega_j|\mathbf{x})$ is maximum. In practice, the jth expert will provide only an estimate $P_j(\omega_i|\mathbf{x})$ of the true a posteriori class probability $P(\omega_i|\mathbf{x})$. The idea of classifier combination is to obtain a better estimate of the a posteriori class probabilities by combining all of the individual expert estimates and, thus, reducing the classification error. A typical estimator is the averaging estimator $\dot{P}(\omega_i|\mathbf{x}) = \frac{1}{N}\sum_{j=1}^{N} P_j(\omega_i|\mathbf{x})$, where $\dot{P}(\omega_i|\mathbf{x})$ is the combined estimate based on N observations.

Let us denote the error on the jth estimate of the ith class aposteriori probability at point $\mathbf{x}$ as $e_j(\omega_i|\mathbf{x})$ and let the probability distribution of the errors be $p_{ij}[e_j(\omega_i|\mathbf{x})]$. Then, the probability distribution of the unscaled error $e_i(\mathbf{x}) = \sum_{j=1}^{N} e_j(\omega_i|\mathbf{x})$ on the combined estimate will be given by the convolution of the component error densities, i.e.,

$$p(e_i(\mathbf{x})) = \int_{-\infty}^{\infty} \dots \int_{-\infty}^{\infty} p_{i1}(\lambda_1) p_{i2}(\lambda_2 - \lambda_1) \dots p_{iN}(e_i(\mathbf{x}) - \lambda_{N-1}) d\lambda_1 d\lambda_2 \dots d\lambda_{N-1}, \quad (1)$$

where λ_i are dummy variables. The distribution of the scaled error $\epsilon_i(\mathbf{x}) = \frac{1}{N} e_i(\mathbf{x})$ is then given by $p(\epsilon_i(\mathbf{x})) = p(\frac{1}{N} e_i(\mathbf{x}))$.

In order to investigate the effect of classifier combination, suppose the a posteriori probability of class ω_s is maximum, giving the local Bayes error $e_B(\mathbf{x}) = 1 - \max_{i=1}^{2} P(\omega_i|\mathbf{x})$. However, our classifiers only estimate these a posteriori class probabilities and the associated estimation errors may result in suboptimal decisions and, consequently, in an additional classification error. In order to quantify this additional error, we have to establish what the probability is for the recognition system to make a suboptimal decision. This situation will occur when the a posteriori class probability estimates for the other class become maximum. Let us derive the probability of the event occurring for a single expert j for class $\omega_i,\ i \neq s$, i.e., when $P_j(\omega_i|\mathbf{x}) - P_j(\omega_s|\mathbf{x}) > 0$. Note that the left hand side of of the inequality can be expressed as

- *J. Kittler is with the Center for Vision, Speech, and Signal Processing, School of Electronics, Computing, and Mathematics, University of Surrey, Guildford, Surrey GU2 7XH, UK. E-mail: j.kittler@eim.surrey.ac.uk.*
- *F.M. Alkoot is with the Telecommunications and Navigation Institute, Public Authority for Applied Education and Training, Shuwaikh, Kuwait. E-mail: f_alkoot@yahoo.com.*

Manuscript received 8 Oct. 2001; revised 20 Mar. 2002; accepted 16 May 2002. Recommended for acceptance by V. Govindaraju.
For information on obtaining reprints of this article, please send e-mail to: tpami@computer.org, and reference IEEECS Log Number 115149.

$$P(\omega_i|\mathbf{x}) - P(\omega_s|\mathbf{x}) + e_j(\omega_i|\mathbf{x}) - e_j(\omega_s|\mathbf{x}) > 0. \quad (2)$$

Equation (2) defines a constraint for the two estimation errors $e_j(\omega_k|\mathbf{x}), k = 1, 2$ as

$$e_j(\omega_i|\mathbf{x}) - e_j(\omega_s|\mathbf{x}) > P(\omega_s|\mathbf{x}) - P(\omega_i|\mathbf{x}), \quad (3)$$

which, in a two class case, will be satisfied if

$$2e_j(\omega_i|\mathbf{x}) > P(\omega_s|\mathbf{x}) - P(\omega_i|\mathbf{x}).$$

The probability $e_A(\mathbf{x})$ of this event occurring will be given by the integral of the error distribution under the tail defined by the margin $\Delta P_{si}(\mathbf{x}) = P(\omega_s|\mathbf{x}) - P(\omega_i|\mathbf{x})$, i.e.,

$$e_A(\mathbf{x}) = \int_{\Delta P_{si}(\mathbf{x})}^{\infty} p_{ij}[2e_j(\omega_i|\mathbf{x})]de_j(\omega_i|\mathbf{x}). \quad (4)$$

In contrast, after fusion, the probability of the additional error is given by

$$e_S(\mathbf{x}) = \int_{\Delta P_{si}(\mathbf{x})}^{\infty} p[2\epsilon_i(\mathbf{x})]d\epsilon_i(\mathbf{x}). \quad (5)$$

Now, how do these labeling errors translate to classification error probabilities? We know that, for the Bayes minimum error decision rule, the error probability at point $\mathbf{x}$ will be $e_B(\mathbf{x})$. If our pseudo-Bayesian decision rule, i.e., the rule that assigns patterns according to the maximum estimated a posteriori class probability, deviates from the Bayesian rule with probability $e_A(\mathbf{x})$, the local error of the decision rule will be given by $\alpha(\mathbf{x}) = e_B(\mathbf{x})[1 - e_A(\mathbf{x})] + e_A(\mathbf{x})[1 - e_B(\mathbf{x})]$, which simplifies to

$$\alpha(\mathbf{x}) = e_B(\mathbf{x}) + e_A(\mathbf{x})[1 - 2e_B(\mathbf{x})] = e_B(\mathbf{x}) + e_A(\mathbf{x})|\Delta P_{si}(\mathbf{x})| \quad (6)$$

as $[(1 - e_B(\mathbf{x})) - e_B(\mathbf{x})]$ is the absolute value of the margin between the two a posteriori class probabilities. For the multiple classifier system which averages the expert outputs, the classification error probability is

$$\beta(\mathbf{x}) = e_B(\mathbf{x}) + e_S(\mathbf{x})|\Delta P_{si}(\mathbf{x})|. \quad (7)$$

Thus, for a multiple classifier system to achieve a better performance, the labeling error after fusion, $e_S(\mathbf{x})$, should be smaller than the labeling error, $e_A(\mathbf{x})$, of a single expert.

Let us now consider fusion by voting. In this strategy, all single expert decisions are hardened and, therefore, each expert will make suboptimal decisions with probability $e_A(\mathbf{x})$. When combined by voting for the most representative class, the probability distribution of k decisions, among a pool of N, being suboptimal is given by the binomial distribution. A switch of labels will occur whenever the majority of individual expert decisions is suboptimal. Assuming the error distributions are identical, this will happen with probability

$$e_V(\mathbf{x}) = \sum_{k=\lfloor \frac{N}{2}+1 \rfloor}^{N} \binom{N}{k} e_A^k(\mathbf{x})[1 - e_A(\mathbf{x})]^{N-k}. \quad (8)$$

Provided $e_A(\mathbf{x}) < 0.5$, this probability will decrease with increasing N. After fusion by Vote, the error probability of the multiple classifier will then be

$$\gamma(\mathbf{x}) = e_B(\mathbf{x}) + e_V(\mathbf{x})|\Delta P_{si}(\mathbf{x})|. \quad (9)$$

The relationship between the label switching errors $\beta(\mathbf{x})$ and $\gamma(\mathbf{x})$ is illustrated in Fig. 1 for Sum and Vote combination strategies for normally distributed estimation errors with different values of $\sigma(\mathbf{x})$ and $N = 3$.

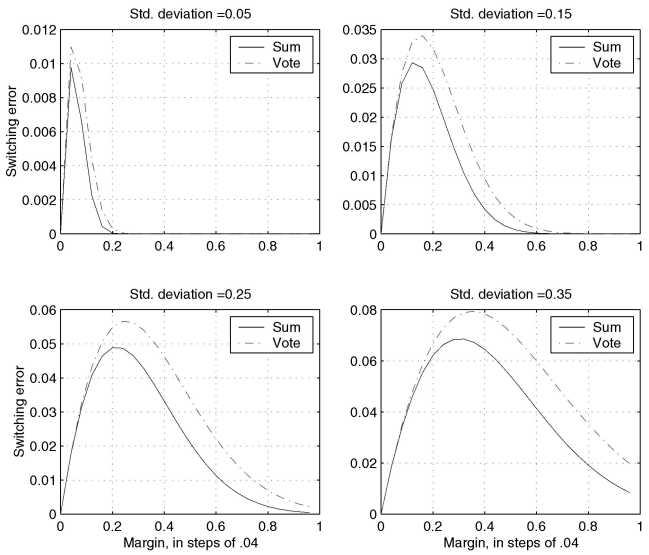

Fig. 1. Sum and Vote switching error for normally distributed estimation errors with different values of $\sigma(\mathbf{x})$ using three experts.

3 Relationship of Sum and Vote

Under the assumption of estimation errors being independent with equal variance $\sigma^2(\mathbf{x})$, the variance of the error distribution for the combined estimate will be $\dot{\sigma}^2(\mathbf{x}) = \frac{\sigma^2(\mathbf{x})}{N}$. Let us assume that the error distributions $p_{ij}[e_j(\omega_i|\mathbf{x})]$ are Gaussian. Then, the distribution of the difference of the two errors with equal magnitude but opposite sign will also be Gaussian with four times as large variance. The probability of constraint (3) being satisfied is given by the area under the Gaussian tail with a cut off point at $P(\omega_s|\mathbf{x}) - P(\omega_i|\mathbf{x})$. More specifically, this probability, $e_A(\mathbf{x})$, is given by

$$e_A(\mathbf{x}) = 1 - \frac{1}{2\sqrt{2\pi}\dot{\sigma}} \int_0^{\Delta P_{si}(\mathbf{x})} \exp^{-\frac{1}{2}\frac{\gamma^2}{4\sigma^2}} d\gamma. \quad (10)$$

In order to compare the performance gains of the Sum and Vote fusion under the Gaussian assumption, we have designed a simulation experiment involving N experts, each estimating the same a posteriori probability $P(\omega_i|\mathbf{x})\ \ i = 1, 2$. Estimation errors are simulated by perturbing the target probability $P(\omega_i|\mathbf{x})$ with statistically independent errors drawn from a Gaussian distribution with a zero mean and standard deviation $\sigma(\mathbf{x})$. We have chosen the a posteriori probability of class ω_1 to always be greater than 0.5. The decision margin $\Delta P_{12}(\mathbf{x})$ is given by $2P(\omega_1|\mathbf{x}) - 1$. The Bayesian decision rule assigns all the test patterns to class ω_1. For each test sample, the expert outputs are combined using the Sum rule and the resulting value compared against the decision threshold of 0.5.

Similarly, the decision errors of the majority vote are estimated by converting the expert outputs into class labels using the pseudo Bayesian decision rule and then counting the support for each class among the N labels. The label of the winning class is then checked against the identity of the test pattern and any errors recorded. The results are averaged over 500 experiments for each combination of $P(\omega_1|\mathbf{x})$ and $\sigma(\mathbf{x})$, the parameters of the simulation experiment.

Typical results showing the additional error incurred are plotted as a function of the number of experts N in Fig. 2. The theoretical values predicted by (5) and (8) are also plotted for comparison. The experimental results mirror closely the theoretically predicted behavior, i.e., Sum being superior to Vote.

While, under the Gaussian assumption, the Sum rule always outperforms Vote, it is pertinent to ask whether this relationship holds for other distributions. Intuitively, if the error distribution has heavy tails, it is easy to see that fusion by Sum will not result in improvement until the probability mass in the tail of $p_{ij}[e_j(\omega_i|\mathbf{x})]$

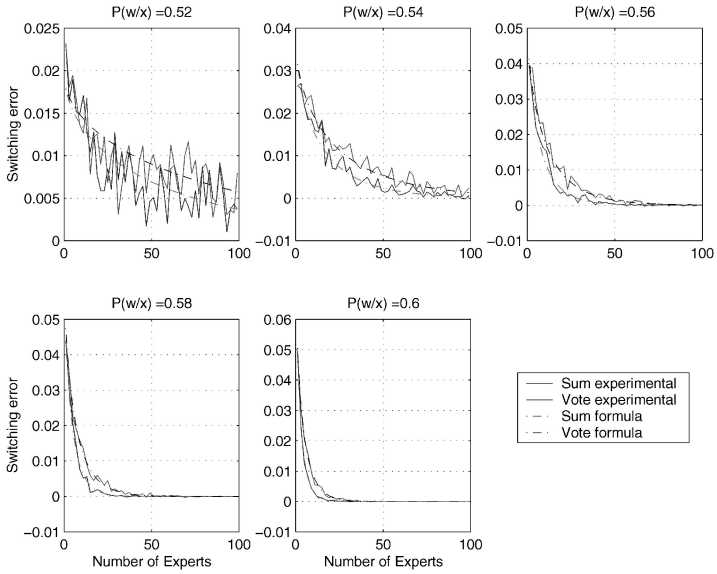

Fig. 2. Comparison of experimental Sum and Vote switching errors with theoretical predictions, for $\sigma(\mathbf{x}) = 0.15$.

moves within the margin $\Delta P_{12}(\mathbf{x})$. In order to gain better understanding of the situation, let us consider a specific example with the error distribution $p_{ij}[e_j(\omega_i|\mathbf{x})]$ being defined as a mixture of three Dirac delta functions with the weights and positions shown in Fig. 3. Using the convolution integral in (1) and substituting into (5), we can derive the probability, $e_S(\mathbf{x})$ of the decision rule being suboptimal for a given margin $\Delta P_{12}(\mathbf{x})$. Fig. 4 shows this probability as a function of the number of expert outputs fused. The function has been computed for a range of margins from $\Delta P_{12}(\mathbf{x}) = 0.04$ to $\Delta P_{12}(\mathbf{x}) = 0.2$. The figure shows clearly an oscillating behavior of $e_S(\mathbf{x})$. It is interesting to note that, for small margins, initially (i.e., for a small number of experts), the error probability of the sum combiner has a tendency to grow above the probability of the decision rule being suboptimal for a single expert. First, the performance improves when $N = 2$, but as further experts are added, the error builds up as the probability mass shifts from the origin to the periphery by the process of convolution. It is also interesting to note that, for $N = 2$, Vote degrades in performance. However, this is only an artifact of a vote tie not being randomized in the theoretical formula. Once the first line of the probability distribution of the sum of estimation errors falls below the threshold defined by the margin between the two class a posteriori probabilities, the performance dramatically improves. However, by adding further experts, the error build up will start all over again, though it will culminate at a lower value than at the previous peak. We can see that, for instance, for $\Delta P_{12}(\mathbf{x}) = 0.04$, the benefits from fusion by the sum rule will be very poor and there may be a wide range of N for which fusion would result in performance deterioration.

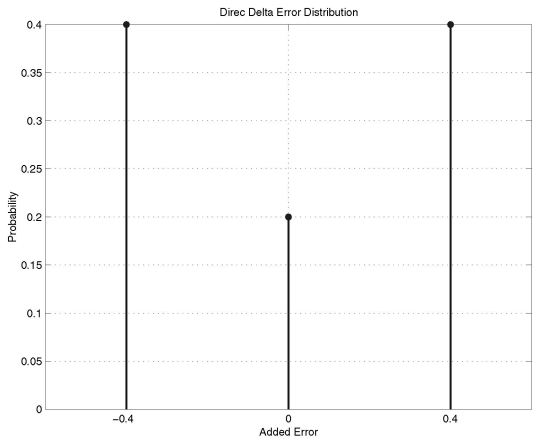

Fig. 3. Dirac delta error distribution.

Once the margin reaches 0.16, Sum will generally outperform Vote, but there may be specific numbers of experts for which Vote is better than Sum. The same kind of behavior is demonstrated in Fig. 5, where the position of the Dirac delta components of the error distribution offset from the origin is at $\pm[1 - P(\omega_1|\mathbf{x})]$, which shows the additional effect of sampling the a posteriori class probability distribution inherent in the simulation approach.

In contrast, the corresponding probability $e_V(\mathbf{x})$, given for the majority vote by (8), diminishes monotonically (also in an oscillating fashion) with the increasing number of experts. Thus, there are situations where Vote outperforms Sum. Most importantly, this is likely to happen close to the decision boundary where the margins are small.

4 Real Data Experiments

In this section, we shall compare Sum and Vote on real data. Real pattern classification problems differ from idealized situations in many different ways. First of all, the point-wise analysis performed in Sections 2 and 3 is impossible as we do not have enough data at each and every single point of the expert output space. Second, in realistic scenarios, the ground truth, at best, is known only coarsely, in terms of class labels rather than true a posteriori probabilities. Third, expert outputs are likely to be correlated. Moreover, it is unlikely that each expert would be estimating the same aposteriori probability functions. Nevertheless, Sum and Vote are useful practical fusion rules and it is interesting to know how they compare when we depart significantly from the underlying assumptions.

As a vehicle for our experimental study, we consider the problem of personal identity verification using face and voice biometrics extracted from the multimedia data in the XM2VTS database [3]. The XM2VTS database is a multimodal database consisting of face images, video sequences, and speech recordings taken of 295 subjects at one month intervals. The database contains four sessions. The Lausanne protocol splits randomly all 295 subjects into 200 clients, 25 evaluation impostors, and 70 test impostors. Our experiments are

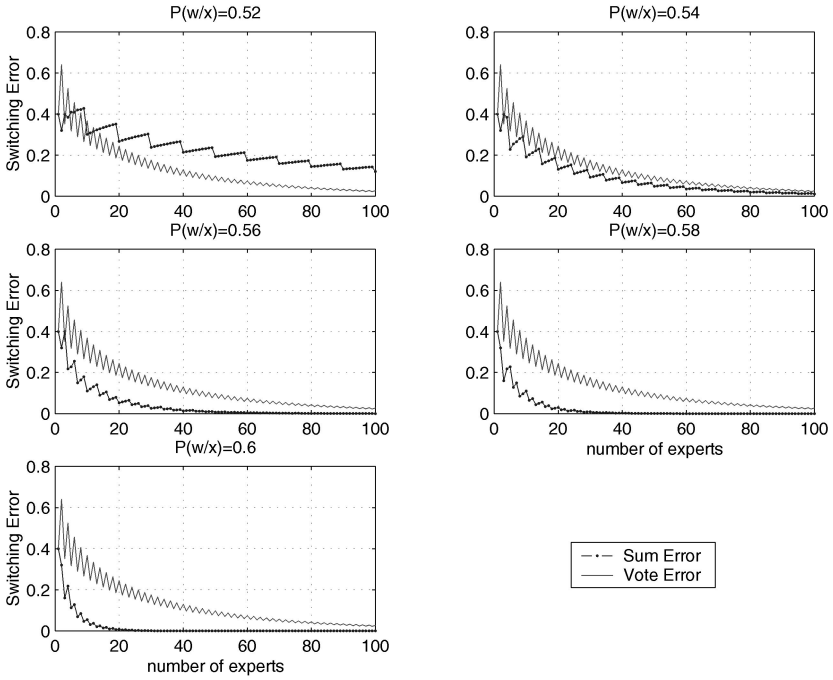

Fig. 4. Theoretical switching error of Sum and Vote in the presence of delta noise at 0.2.

based on evaluation protocol in configuration I for which the evaluation set contains 600 client shots (200 clients × three shots) and 40,000 imposter cases (25 impostors × 8 shots × 200 clients). The test set contains 400 client shots (200 clients × two shots) and 112,000 imposter cases (70 impostors × eight shots × 200 clients).

Scores from eight different experts are used as our single expert outputs that we need to combine. FACE2 and FACE4 are two of the experts designed at the University of Surrey which confirm or reject the claimed identity using face biometrics. SPEECH2 and SPEECH3 experts designed at IDIAP, Switzerland, base the identity on the speaker's voice characteristics. Experts numbered five to seven are from the Aristotle University of Thessaloniki and are based on elastic graph matching. The methods differ in the internal threshold settings which, respectively, favor low rejection rates, low false acceptance rates, and equal error rates. SydneyCI is the eighth expert, from the University of Sydney, based on fractal image coding. The expert scores on the XM2VTS database are available from [3]. The database and the eight experts used in our experiments are described in [10].

Using the combiner performance on the evaluation set, we select three different threshold values:

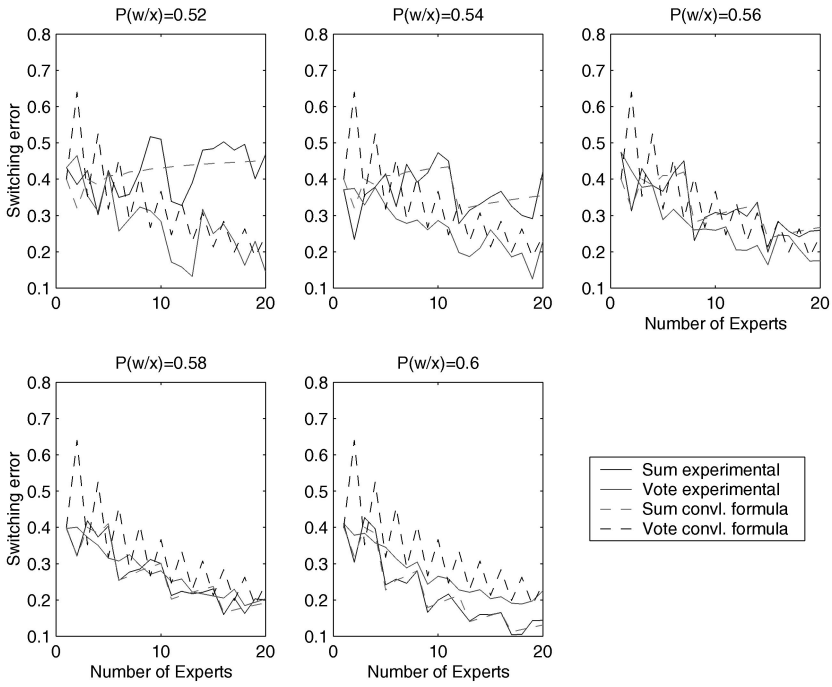

Fig. 5. Sum and Vote switching error: a comparison of single class experimental results and theoretical predictions for delta noise located at (1-p) for up to 20 experts.

TABLE 1
Average Correct Verification Rates of the Individual Experts

Set type	FR=0	FA=0	EER	FR=0	FA=0	EER
	FACE2 = expert number 1			FaceautEER = expert number 5		
Eval.	0.5183	0.8667	0.9434	0.5039	0.8250	0.9201
Test	0.5165	0.8625	0.9282	0.5048	0.7487	0.9294
	FACE4 = expert number 2			FaceautFA = expert number 6		
Eval.	0.5945	**0.9192**	0.9650	0.5045	0.8125	0.9184
Test	0.5942	**0.9275**	0.9690	0.5048	0.7312	0.9249
	SPEECH2 = expert number 3			FaceautFR = expert number 7		
Eval.	0.9935	0.5575	0.9938	0.5059	0.8058	0.9183
Test	0.9669	0.5237	0.9580	0.5067	0.7312	0.9269
	SPEECH3 = expert number 4			SydneyCI = expert number 8		
Eval.	**0.9999**	0.5000	**0.9991**	0.5281	0.6467	0.8708
Test	**0.9926**	0.5000	**0.9926**	0.5301	0.5937	0.8706

$$T_{FAE=0} = \arg\min_T(FRE|FAE = 0)$$
$$T_{ERR} = T_{FAE=FRE} = (T|FAE = FRE)$$
$$T_{FRE=0} = \arg\min_T(FAE|FRE = 0),$$

where FAE and FRE are the false acceptance and false rejection error rates using the evaluation set. Using each of these threshold values, we measure the average correct verification rate. All of the rates are measured on both the evaluation set and the test set.

The strength of the individual experts can be gleaned from the single expert results in Table 1. We note that the SPEECH3 expert is the best expert for T_{EER} and $T_{FRE=0}$, while FACE4 is the best for $T_{FAE=0}$. The test set results confirm the performance on the evaluation set. It is also interesting to note that the performance of the individually best expert can be as much as an order of magnitude better than that of the second best.

When fusing the available experts, we have the choice of combining any subset of them. The number of experts combined, N, could range from two up to eight experts. For each value of N and each threshold type, we find the set of experts that performs best on the evaluation set. We notice that, although FACE2 has a lower performance than FACE4, it is preferred by both combiners, Sum and Vote. It is consistently selected along with the two speech experts. Actually, Sum selected FACE2 and SPEECH4 for $T_{FRE=0}$ and T_{EER} for all values of N. This leads us to conclude that we should not be looking for the best single expert to combine, but the most complementary ones.

The fusion results obtained with the best combinations of experts for the Sum and Vote are shown in Table 3. Overall, we can see that, as predicted theoretically, initially, for a small number of experts ($N = 2$), Sum is the best. As the number of experts increases, Vote is better than Sum, but, for $N > 5$, Sum outperforms Vote. However, in contrast to the results of the simulation experiments, in this parameter range, we fail to observe a monotonic improvement of both fusion strategies. This is due to the fact that our experts have unequal strengths and, therefore, one of the assumptions made is not satisfied. Although we reduce the variance of the probability distribution of the fused decision being suboptimal by including more experts, the Bayes error rate decreases and so does the overall performance of the multiple classifier system.

Not surprisingly, the SPEECH3 expert, the best single expert, is always selected as one of the experts to be fused. For the Sum rule, the performance on the evaluation set indicates that the combination of $2 - 4$ experts for $T_{FRE=0}$ and T_{EER} yield the results which are flat as indicated in Table 2. For $T_{FAE=0}$, the best results with Sum in this range of N slightly oscillate. Thus, the Sum rule appears to exhibit the same oscillatory behavior that we noted in the simulation experiment with a nonGaussian probability distribution $p_{ij}[e_j(\omega_i|x)]$. This would suggest that error distribution has relatively heavy tails, probably due to one of the speech experts being based on HMM.

From the test set results shown in Table 3, we find that the best performance for $T_{FAE=0}$ using Sum is delivered when N = 2, as was suggested by the evaluation set. Similarly, for $T_{FRE=0}$ and T_{EER},

TABLE 2
The Classification Rates of the Best Mixtures of Experts Obtained on the Evaluation Set

No. of Experts	FR=0 impos rate	FA=0 client rate	EER average rate	FR=0 impos rate	FA=0 client rate	EER average rate
	Sum			Vote		
2	99.9900	**99.8333**	99.9117	99.9975	99.8333	99.9154
	14	**34**	14	34	34	34
3	99.9900	98.3333	99.9117	**99.9975**	**99.8333**	**99.9154**
	134	148	134	**134**	**134**	**134**
4	**99.9900**	99.0000	**99.9117**	99.9675	99.6667	99.9004
	1348	1348	**1348**	1348	1348	1348
5	99.9875	98.8333	99.9104	99.9675	99.6667	99.9004
	12348	12348	12348	12348	12348	12348
6	99.9375	98.5000	99.8867	99.2925	98.5000	99.5113
	123468	123468	123468	123478	123478	123478
7	99.3625	97.6667	99.8329	41.6325	98.5000	98.8329
	1234568	1234568	1234568	1234678	1234678	1234678

The expert identity in each mixture is indicated by their ID number below each rate. The best number of experts for each threshhold type and combiner is indicated in bold. To obtain expert names from their ID, refer to Table 1.

TABLE 3
The Test Set Classification Rates of the Best Mixtures of Experts Selected on the Evaluation Set

No. of experts	FR=0	FA=0	EER	FR=0	FA=0	EER
	Sum			Vote		
2	0.9930	**0.9962**	0.9931	0.9759	0.9722	0.9734
3	0.9938	0.9945	0.9938	**0.9934**	**0.9922**	**0.9921**
4	**0.9944**	0.9893	**0.9944**	0.9955	0.9922	0.9955
5	0.9943	0.9932	0.9944	0.9955	0.9922	0.9955
6	0.9942	0.9800	0.9942	0.9910	0.9910	0.9929
7	0.9916	0.9874	0.9940	0.7116	0.9947	0.9917
8	0.9406	0.9860	0.9918	0.5322	0.9786	0.9718

the best performance is when N = 4, which is in agreement with the best performance on the evaluation set.

For the vote rule, the evaluation set yields the highest performance rate when N = 2 or N = 3. Recall from the synthetic data experiments that we should normally avoid using an even number of experts for the vote. Thus, we should select as the best combination experts, FACE2, SPEECH2, and SPEECH3, indicated in bold in Table 2. Based on the test set, we find that Vote fails to yield the best performance at N = 2 or 3. Its performance peaks when N = 4 and 5, which would not be considered, given the information provided by the evaluation set.

Note, in Table 2, that the performance of the sum rule does not improve as the number of experts increases. This would suggest that Vote might be preferable to Sum. However, as Vote also fails to improve as the number of experts increases, we know from the theoretical expectations that we are most likely combining experts of unequal strength. In these circumstances, a more conservative option is to adopt the Sum rule, as the relationship of its results on the evaluation and test sets appears to be less volatile. This strategy, based on the results of this paper, gives quite good performance on the test set, almost matching the absolute best performance achievable a posteriori.

5 Discussion and Conclusion

We studied the relationship of the Sum and Vote fusion strategies and showed, analytically, that, for Gaussian distributions of estimation errors, the Sum fusion strategy will always outperform the majority vote. However, for heavy tail distributions, the majority vote may give better results than Sum. This may happen when the margin between the two a posteriori class probabilities is small. However, even if the margin is reasonable, Vote may be superior to Sum when the number of experts is small. This behavior may explain the conflicting observations made by a number of researchers about the two rules.

The above ideal behavior of the two rules applies only under the assumption that the expert errors are independent and identically distributed. Moreover, the target a posteriori class probability estimated by each expert must be the same. In practice, none of the assumptions are likely to hold exactly. The effect of the first two not being satisfied is likely to be reflected in slower rates of performance improvement than those predicted by the theory as the number of experts increases.

The implications of the third assumption being invalid are much more serious. Here, as we are adding more experts to reduce variance, we may be including experts which inject additional uncertainty in the input of the multiple classifier system. Thus, any gains in reduced variance will have to be weighed against potential losses due to increased uncertainty. This raises a serious methodological question of how to decide when the fusion of more experts starts being counterproductive, which is currently being investigated.

We then confirmed our theoretical predictions by experiments on synthetic data. Experiments on real data supported the general findings, but also showed the effect of the usual assumptions of conditional independence, identical error distributions, and common target outputs of the experts not being fully satisfied.

Acknowledgments

This work was supported by the Engineering and Physical Sciences Research Council under Grant GR/M61320 and European Project Banca. F.M. Alkoot would like to thank the Islamic Development Bank and the Public Authority for Applied Education and Training for financially supporting his education toward his PhD degree.

References

[1] F.M. Alkoot and J. Kittler, "Experimental Evaluation of Expert Fusion Strategies," *Pattern Recognition Letters*, vol. 20, no. 11, pp. 11-13, 1999.
[2] L. Breiman, "Bagging Predictors," *Machine Learning*, vol. 24, pp. 123-140, 1996.
[3] The Extended M2VTS Database, http://www.ee.surrey.ac.uk/Research/VSSP/xm2vtsdb/, 2002.
[4] T. Dietterich, "An Experimental Comparison of Three Methods for Constructing Ensembles of Decision Trees: Bagging, Boosting, and Randomization," *Machine Learning*, pp. 1-22, 1998.
[5] R.P.W. Duin and D.M.J. Tax, "Experiments With Classifier Combining Rules," *Multiple Classifier Systems*, J. Kittler and F. Roli, eds., pp. 16-29, Springer, 2000.
[6] L.K. Hansen and P. Salamon, "Neural Network Ensembles," *IEEE Trans. Pattern Analysis and Machine Intelligence*, vol. 12, no. 10, pp. 993-1001, Oct. 1990.
[7] T.K. Ho, J.J. Hull, and S.N. Srihari, "Decision Combination in Multiple Classifier Systems," *IEEE Trans. Pattern Analysis and Machine Intelligence*, vol. 16, no. 1, pp. 66-75, Jan. 1994.
[8] 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, Mar. 1998.
[9] L. Lam and C. Suen, "Application of Majority Voting to Pattern Recognition: An Analysis of Its Behaviour and Performance," *IEEE Trans. Systems, Man, and Cybernetics, Part A: Systems and Humans*, vol. 27, no. 5, pp. 553-568, 1997.
[10] J. Matas, M. Hamouz, K. Jonsson, J. Kittler, Y. Li, C. Kotropoulos, A. Tefas, I. Pitas, T. Tan, H. Yan, F. Smeraldi, J. Bigun, N. Capdevielle, W. Gerstner, S. Ben-Yacoub, Y. Abdeljaoued, and E. Mayoraz, "Comparison of Face Verification Results on the XM2VTS Database," *Proc. Int'l Conf. Pattern Recognition*, 2000.
[11] A.F.R. Rahman and M.C. Fairhurst, "Enhancing Multiple Expert Decision Combination Strategies through Exploitation of A Priori Information Sources," *Proc. Vision Image and Signal Processing*, vol. 146, no. 1, pp. 40-49, 1999.
[12] C. Suen, R. Legault, C. Nadal, M. Cheriet, and L. Lam, "Building a New Generation of Handwriting Recognition Systems," *Pattern Recognition Letters*, vol. 14, pp. 303-315, 1993.
[13] D.M.J. Tax, M. van Breukelen, R.P.W. Duin, and J. Kittler, "Combining Multiple Classifiers by Averaging or by Multiplying," *Pattern Recognition*, vol. 33, no. 9, pp. 1475-1485, 2000.
[14] K. Tumer and J. Ghosh, "Analysis of Decision Boundaries in Linearly Combined Neural Classifiers," *Pattern Recognition*, vol. 29, no. 2, pp. 341-348, 1996.
[15] L. Xu, A. Krzyzak, and C.Y. Suen, "Methods of Combining Multiple Classifiers and Their Applications to Handwriting Recognition," *IEEE Trans. Systems, Man, and Cybernetics*, vol. 22, no. 3, pp. 418-435, 1992.