Investigating The Fusion of Classifiers Designed Under Different Bayes Errors

Fuad M. Alkoot
Telecommunication and Navigation Institute, PAAET
P.O.Box 4575, Alsalmia, Kuwait
f_alkoot@yahoo.com
and
Josef Kittler
CVSSP, University of Surrey
Guildford, GU2 7XH

Abstract

We investigate a number of parameters commonly affecting the design of a multiple classifier system in order to find when fusing is most beneficial. We extend our previous investigation to the case where unequal classifiers are combined. Results indicate that Sum is not affected by this parameter, however, Vote degrades when a weaker classifier is introduced in the combining system. This is more obvious when estimation error with uniform distribution exists.

Key words: Fusion, bayes error, classifier combination, Vote, Sum.

1 Introduction

In the past decade, the use of classifier fusion to improve classification accuracy has become increasingly popular [2, 3, 4, 5, 6, 9, 10, 12, 14, 15, 16]. Classifier combination has been attracting considerable attention because of its potential to ameliorate the performance of pattern recognition systems. The basic idea is to solve each pattern recognition problem by designing a number of classification systems and then combining the designs in some way to achieve reduced recognition error rates. The fusion process may operate on the soft outputs of the individual experts, or it may involve combining the hard decisions of the experts. The literature on classifier combination grows rapidly and by now includes hundreds of articles.

The study is based on the fact that pattern recognition systems commonly commit errors that may exceed the bayes error. Here we investigate the design of classifier combination systems involving classifiers designed with different underlaying bayes errors. Additionally it is known that the actual error of a pattern classifier is constituted of two components: the bayes error and an additional error. In [7] we showed the additional error to be the ”switching error” and the margin between the a posteriori probability of the class the pattern optimally belongs to and the class resulting from the incorrect labeling assinment.

The switching error originates from the combined effect of fitting and estimation errors associated with the training proccess used to design the classifiers. The fitting error originates from the design decisions regarding the underlying model of the classification process and its complexity (number of free parameters). The estimation error reflects the fact that the training set on which the design is based is finite and therefore will not adequately represent the class populations.

As the switching error is stochastic depending on the randomness of the sampling and estimation processes, it can be reduced by classifier fusion. In a combining system, when each classifier commits a different switching error, the rest of the classifiers may compensate the effect of that error. Thus the overall system error may be reduced to the bayes error. For example when Sum [8, 1, 11] is used to combine several classifiers, the estimation error variance will be reduced [13], leading to the reduction in switching error.

In our investigations we study the circumstances where there is a merit in combining classifiers of unequal strength, i.e. classifiers designed with different underlaying bayes error.

This should ultimately provide us with a methodology which can be used for determining if there is a merit in combining weak classifiers. The answers depend on the distribution of errors on the probability estimates. On the margin between class a posteriori probabilities. And the difference between the bayes errors of the classifiers being fused. These parameters indicate the problem is too complex to draw any general conclusion. In this paper we tackle the third parameter and show the effect of the difference between bayes errors. By simulation studies we shall derive the conditions under which the performance of two classifiers will improve. These results will then be extended to the case of fusing multiple classifiers. Results indicate that Vote degrades as weaker classifiers are used. The use of two equal classifiers yields better results than using a stronger and a weaker classifier.

The paper is organized as follows. In Section 2 we shall introduce the necessary notation and present the mathematical preliminaries. In Section 3 the experimental methodology is introduced. The results of modeling are presented in Section 4 and their implication discussed in Section 5. Finally the paper is drawn to conclusion in Section 6.

2 Mathematical Preliminaries

Let us denote the aposteriori probability of class ω_i given observation (pattern) x by $P(\omega_i|x)$. Suppose class ω_s satisfies

$$P(\omega_s|x) = \max_i^m P(\omega_i|x) \tag{1}$$

where m denotes the number of classes. Thus the Bayes optimal decision would be to assign pattern x to class ω_s . Let class ω_j satisfy

$$P(\omega_j|x) = \max_{i=1,i\neq s}^m P(\omega_i|x) \tag{2}$$

Thus in the presence of estimation errors the most likely suboptimal decision will be to assign pattern x to class ω_j. The probability of the label switching error $e_s(x)$ will depend on the distribution $p(\epsilon_i(x))$ of errors $\epsilon_i(x)$ corrupting the estimate of the ith class aposteriori probability. It has been shown in [1] that the switching error $e_s(x)$ is given by

$$e_s(x) = \int_{\Delta P}^{\inf} p(\epsilon_i(x))dt \tag{3}$$

where $\Delta P(x)$ is the margin between the aposteriori probabilities of the two classes likely to be swapped, i.e.

$$P(x) = P(\omega_s|x) - P(\omega_j|x) \tag{4}$$

The additional error $e_A(x)$ [1] is then given as

$$e_A(x) = e_s(x)\Delta P(x) \tag{5}$$

Assuming that the probability of switching between class ω_s and any other class ω_i, $i \neq j$, is negligible, the actual classifier error e(x) will then be

$$e(x) = e_B(x) + e_s(x)\Delta P(x) \tag{6}$$

Note that in a two class case $\Delta P(x)$ in (4) can be expressed as

$$\Delta P(x) = 1 - 2e_B(x) \tag{7}$$

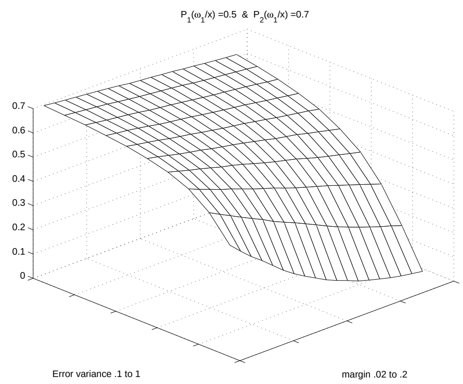
Thus the additional error in (6) can be written as

$$e(x) = e_B(x)[1 - 2e_s(x)] + e_s(x) \tag{8}$$

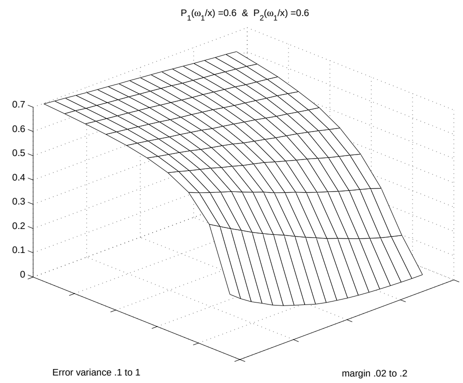
In the multiclass case the margin $\Delta P(x)$ will, in general, be greater than $1 - 2e_B(x)$. However, the above assumption that the switching error between the Bayes optimal decision and any other class ω_i, $i \neq j$ is negligible, implies that $P(\omega_i|x) = 0$, for all , $i \neq s$, j in the equation in (7) will be valid and it will represent the worst case scenario.

3 Experimental Methodology

As shown in Section 2, the actual error of a classifier depends on the Bayes error and on the switching error. The switching error itself is a function of the Bayes error and the probability distribution of the estimation error. We experiment with two estimation error distributions, gaussian and uniform. In our experiments we vary the standard deviation of the normal error distribution from .1 to 1,

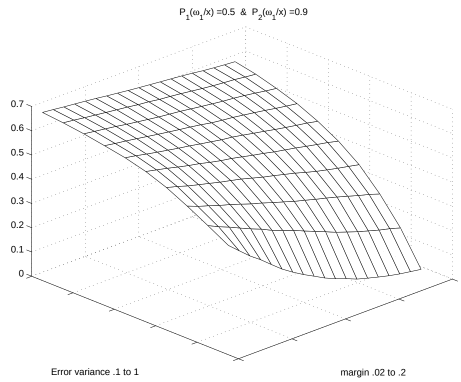


(a)

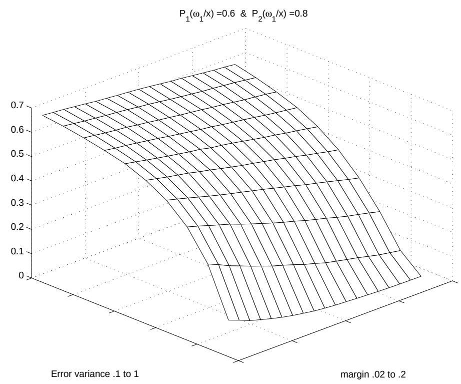


(b)

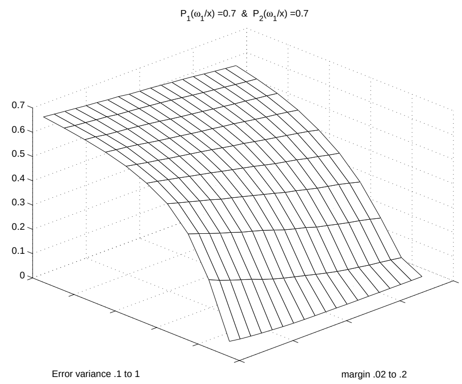
Figure 1: Vote fusion using normal distribution for an average posterior probability of 0.6. Figure (a) is using unequal classifiers while figure (b) is when using equal classifiers



(a)



(b)



(c)

Figure 2: Vote fusion using normal distribution for an average posterior probability of 0.7. Figure (c) is using equal classifiers

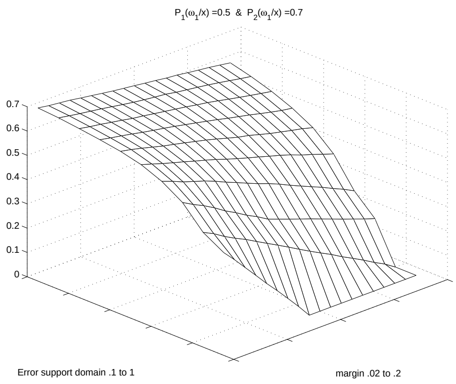
Table 1: True posterior probability values of class 1 for each classifier of the systems under investigation

Case	System	Class 1 post. prob.	
		Expert 1	Expert 2
1	Equal strength	0.6	0.6
	Unequal strenth	0.5	0.7
2	Equal strength	0.7	0.7
	Unequal strenth	0.6	0.8
	Unequal strenth	0.5	0.9
3	Equal strength	0.8	0.8
	Unequal strenth	0.7	0.9

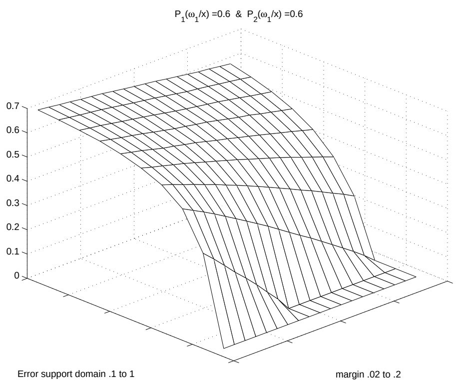
while half of the support domain of the uniform distribution is varied between .1 and 1, at steps of 0.1. Both distributions have a zero mean. In our experiments we only combine two experts. The two experts may have equal strength, i.e. similar bayes error, or unequal strength. In order to compare compatible systems, for all the experiments we select the unequal experts posterior probability such that their average is equal to that of the equal strength experts. For example, when the posterior probability of the equal experts is 0.6 the unequal experts posterior probabilities are set at 0.7 and 0.5. Table 1 shows the different values which we experiment with. An estimation error with the distributions described above is then added to these posterior probabilities. For each true posterior probability of table 1, we vary the margin between the classes by adding Δ to the class 1 posterior probability. Therefore, for experiments involving posterior probability of 0.6 we add .01 up to .2 with increments of 0.01. the second class posterior probability is reduced by the same amount. In all of the figures, the x-axis represent the added Δ which increases the margin between the classes. The y-axis represent increases in the error distribution standard deviation or support domain, while the z-axis represent the degree of classification error.

4 Experimental Results

When fusing experts using Sum we de not notice any difference between the systems incorporating similar experts and the systems with unequal experts. However, using equal experts yields a negligible improvement over unequal experts when an estimation error with uniform distribution exists. When Vote is used the equal expert combiner outperforms the unequal expert combiner. The difference becomes more when estimation error with uniform distribution exists. The difference between the two combiner systems is more obvious at the lower values of the standard deviation or support domain. Results indicate that the system with expert probabilities of 0.6 and 0.8 outperforms the system with 0.5 and 0.9 probabilities when Vote is used. However, if Sum is used both systems perform equally.



(a)



(b)

Figure 3: Vote fusion using uniform distribution for an average posterior probability of 0.6. Figure (a) is using unequal classifiers while figure (b) is when using equal classifiers

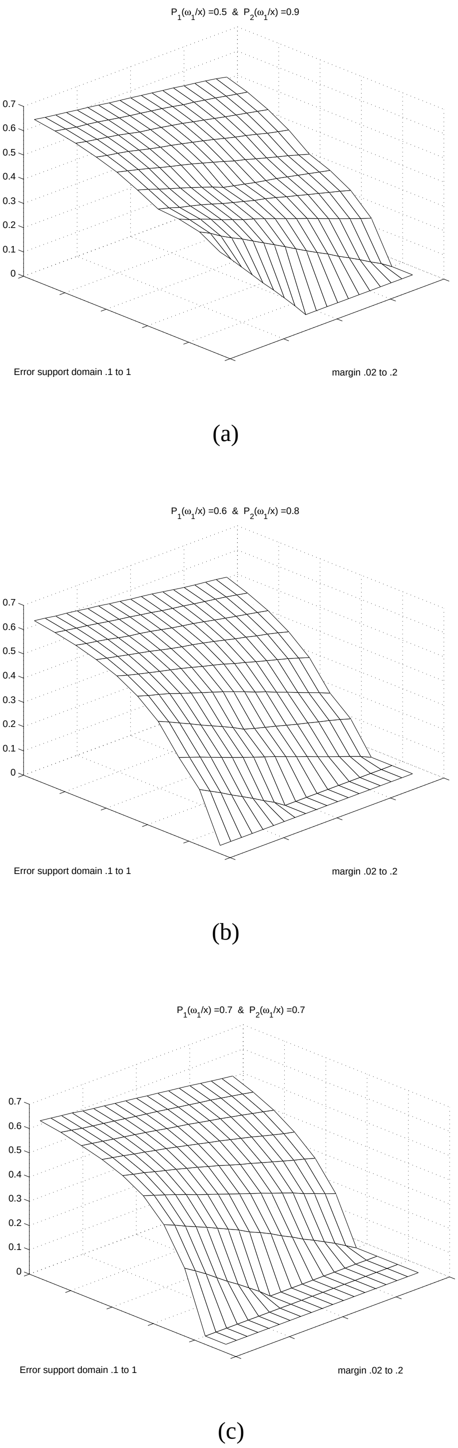


Figure 4: Vote fusion using uniform distribution for an average posterior probability of 0.7. Figure (c) is using equal classifiers

5 Discussion

It is obvious that Sum will not be affected by the types of experts used. That is due to the fact that Sum averages the estimates of the experts, and in our experiments we choose to compare systems containing similar bayes errors on average. Also, note that the added estimation error is randomly added to all experts using similar distributionss with a zero mean. From equation (6) for systems with equal experts we get

$e(x) = 0.6 + 0.2e_s(x)$, however for unequal experts we get

$$e(x) = \frac{(0.7+0.5)+(0.4\times e_s(x))+(0\times e_s(x))}{2} = \frac{0.12+0.4\times e_s(x)}{2} = 0.6 + 0.2 \times e_s(x)$$

Using vote fusion the probability estimates are not fused, however, the decisions or labels are fused. Therefore, when the true class probability falls below 0.5 a fusion classification error may occur. Consequently, the two experts with true probability of 0.6 may commit less error than when one expert has 0.5 bayes error even if the other has a bayes error of 0.3. this can be verified from the figures.

Further investigations are being carried out to find the effect of the different parameters on the design and performance of a multiple classifier system, when unequal classifiers are combined.

6 Conclusion

We investigate the performance of sum and vote fusion strategies when combining classifiers of unequal strength, i.e. different underlying bayes error. This is compared to the case involving equal strength classifiers. Experimental results show that both fusion systems yield similar results when combined using Sum. This is due to the fact that all systems yield similar results if the classifiers are on average equal. However, Vote yields a superior performance when combining equal strength classifiers compared to when combining unequal classifiers. This is attributed to the fact that Vote fuses labels while Sum averages expert outputs before labeling.

References

[1] F. M. Alkoot and J. Kittler. Experimental evaluation of expert fusion strategies. *Pattern Recognition Letters*, 20(11-13):1361–1369, 1999.

[2] F. M. Alkoot and J. Kittler. Multiple expert system design by combined feature selection and probability level fusion. In *Proceedings of the Fusion 2000 conference*, volume II, pages ThC5(9–16), Paris, France, 2000.

[3] T. Dietterich. An experimental comparison of three methods for constructing ensembles of decision trees: Bagging, boosting, and randomisation. *Machine Learning*, pages 1–22, 1998.

[4] M C Fairhurst and A F R Rahman. Generalised approach to the recognition of structurally similar handwritten characters using multiple expert classifiers. *IEE Proc. on Vision, Image and Signal Processing*, 144(1):15–22, 1997.

[5] T.K. Ho, J.J. Hull, and S.N. Srihari. Decision combination in multiple classifier systems. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 16(1):66–75, 1994.

[6] M I Jordan and R A Jacobs. Hierarchical mixture of experts and the em algorithm. *Neural Computation*, 6:181–214, 1994.

[7] J. Kittler and F. M. Alkoot. Relationship of sum and vote fusion strategies. In J. Kittler and F. Roli, editors, *2nd International workshop on multiple classifier systems*, pages 339–348, Cambridge, UK, 2001. Springer.

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

[9] J. Kittler, J. Matas, K. Johnson, and M. RamosSanchez. Combining evidence in personal identity verification systems. *Pattern Recognition Letters*, 18:845–852, 1997.

[10] G. Rogova. Combining the results of several neural network classifiers. *Neural Networks*, 7(5):777–781, 1994.

[11] D. Tax, M. van Breukelen, R. Duin, and J. Kittler. Combining multiple classifiers by averaging or by multiplying. *Pattern Recognition*, 33(9):1475–1485, 2000.

[12] K. Tumer and J. Ghosh. Order statistics combiners for neural classifiers. In *Proceedings of the World Congress on Neural Networks*, volume I, pages 31–34, Washington, DC., 1995.

[13] K. Tumer and J. Ghosh. Analysis of decision boundaries in linearly combined neural classifiers. *Pattern Recognition*, 29(2):341–348, 1996.

[14] D.H. Wolpert. Stacked generalisation. *Neural Networks*, 5(2):241–260, 1992.

[15] K Woods, W P Kegelmeyer, and K Bowyer. Combination of multiple experts using local accuracy estimates. *IEEE Transaction Pattern Analysis and Machine Intelligence*, 19:405–410, 1997.

[16] L. Xu, A. Krzyzak, and C.Y. Suen. Methods of combining multiple classifiers and their applications to handwriting recognition. *IEEE Transaction. SMC*, 22(3):418–435, 1992.