Rejection Versus Error in a Multiple Experts Environment

Louisa Lam[†‡] and Ching Y. Suen[‡]

[†]Department of Mathematics
Hong Kong Institute of Education
10 Lo Ping Road, Tai Po, Hong Kong

[‡]Centre for Pattern Recognition and Machine Intelligence
Concordia University
1455 de Maisonneuve Blvd. West, Montréal, Québec H3G 1M8, Canada.

Abstract. The combination of classifiers has become a very active research area in recent years, and many results have been obtained through various methods. This paper presents some of our theoretical and experimental work in this domain.

1 Introduction

The recognition of handwritten characters and words has been a very active research area in recent years. The intensive and extensive research efforts by many researchers have resulted in a large number of algorithms. However, due to the immense variety of writing styles that can be used by the general population, the correct identification of unconstrained user-independent handwriting by a single algorithm is still a challenging problem. It is also well-known that improvements in performance by a single method becomes much more difficult to achieve beyond certain levels of performance.

At the same time, practical applications demand high levels of accuracy (or reliability), which is especially true for applications in which errors can be very costly, such as the automatic processing of amounts on cheques and payment forms. For these applications, it would be better to have very low error rates even if it involves a certain amount of rejections (and thus lower recognition rates). These are cases in which the algorithm determines that it cannot make a very confident classification, and therefore it rejects the pattern, which can then be processed by other methods or by a human operator.

The combination of these factors (the abundance of algorithms for the task, the difficulty of achieving completely accurate results by a single algorithm, and the highly reliable results required by real-life applications) had initiated a trend to combine several algorithms in order to produce more reliable results. Consequently the combination of multiple algorithms (multiple expert systems) has been a much studied topic in the past several years ([3], [7], [9], [14], and [17] for example). In the process of designing and studying combination methods, different recognition and rejection rates have been obtained.

In this paper, we will discuss some of our theoretical and experimental work on the rejection-error trade-offs involved in combination methods. The classifiers considered may be independent or otherwise, and the methods are applied to handwritten characters as well as other patterns.

2 Combination by Majority Vote

In combining decisions of pattern classifiers, the method used depends on the nature of the output produced by each recognition algorithm. If this output consists of a single assigned class, then many of the combination methods would not be applicable, and one of the most suitable means of arriving at a combined decision would be by majority vote (in which all votes have equal weight). The voting problem has been much studied since its beginning. When applied to pattern recognition, we have observed various aspects of its behavior which have led to many new theoretical results in the trade-offs between error and rejection rates.

2.1 Odd and Even Numbers of Experts

In our experiments, we have combined the classification decisions of 2 to 9 experts by majority vote. These classifiers have been developed by different researchers and applied to different databases of handwritten numerals.

It has been observed that combinations of even numbers of classifiers tend to produce both lower recognition and lower error rates (and higher rejection rate) than that of odd numbers. Adding one classifier to an even number would increase both the recognition and error rates, while adding this to an odd number would decrease both rates.

Of course, increases in both the recognition and error rates are accompanied by a decrease in the rejection rate, and vice versa. Observing this trade-off between the rejection and error rates has led us to consider the problem theoretically [8], and we established that these results are true regardless of the performances of the individual classifiers and whether they are independent or not.

This is due to the fact that increasing the number of experts from an odd number n to an even number $n+1$ can only change certain decisions to tied votes, resulting in indecisions or rejections. The votes which can be changed this way are the ones in which the initial voting had a majority of only one vote. Similarly, when we add one vote to an even number of votes, the added vote would have the effect of breaking certain ties, thus decreasing rejections by changing them to recognitions and errors. This pattern would always be true, even though the magnitude of the trade-offs would depend on the performance of the particular classifiers.

2.2 Increasing by Two Votes

When we add two votes to an even (or odd) number of votes as repeated additions of one vote, it is not clear what the net effect of the two additions would be, since the second addition appears to reverse the trend of the first. For this reason, we have to examine the results when the two votes are added together to an existing group of votes.

By assuming independence of the votes, we established theoretically [8] that the results depend on the classical entity of odds ratio, where the odds ratio r_i of vote i is defined as

$$r_i = \frac{probability(correct)}{1 - probability(correct)}. \tag{1}$$

If the original n votes have odds ratios r_i for $i = 1, ..., n$, and the new votes have odds ratios s_1 and s_2, then adding the new votes would increase the combined recognition rate if $s_1 s_2 > r_i$ for all i.

Moreover, adding two votes to an even number would be more effective in increasing the recognition rate than adding the votes to an odd number, given similar levels of performance. However, if reducing the error rate is the more important objective, then adding two votes to an odd number would be more effective.

These theoretical results can be observed in Table 1, where results of combining up to nine classifiers are shown. Eight of the classifiers were developed by researchers at CEDAR, State University of New York at Buffalo [10], and the other one at the University of Toronto [15]. The classifiers were applied to the BS Database from CEDAR [1], which was obtained from US mailpieces. These results were combined in ascending order, which means that the weakest two (in terms of recognition rate) classifiers were combined first, then better performing classifiers were added in succession.

Table 1. Results of combining classifiers in ascending order

# Classifiers	2	3	4	5	6	7	8	9
Recognition %	91.52	94.65	93.77	95.39	94.91	96.02	95.72	96.46
Error %	2.55	3.69	2.03	2.43	1.51	2.03	1.40	1.59
Rejection %	5.94	1.66	4.21	2.18	3.58	1.95	2.88	1.95
Reliability %	97.29	96.25	97.88	97.51	98.43	97.93	98.56	98.38

From Table 1, we can see exactly the zigzag effect created by additions of one vote. In addition, it is clear that addition of two votes to an even number is more effective in increasing the recognition rate than adding two votes to an odd number, while the latter is more effective in reducing the error rate. These changes are summarized in Table 2.

Table 2. Effect of adding two classifiers in ascending order

Action	2 to 4	3 to 5	4 to 6	5 to 7	6 to 8	7 to 9
Increase in Recognition %	2.25	0.74	1.14	0.63	0.81	0.44
Decrease in Error %	0.52	1.26	0.52	0.4	0.11	0.44

The results of combining these classifiers in ascending and descending orders are illustrated in Fig. 1. In this figure, the results obtained from combining classifiers in ascending order of performance are denoted by dotted lines, and the number of classifiers is represented with the symbol '#' following the number. Understandably, combining classifiers in descending order (which means the best classifiers are combined first) give better initial results, as indicated by the solid lines in this figure. However, the same zigzag pattern is obtained from the addition of one vote. Eventually, the addition of two votes in descending order to six and seven votes fail to produce better results, showing that the successively weaker votes can no longer satisfy the condition required for improvement in the combined result.

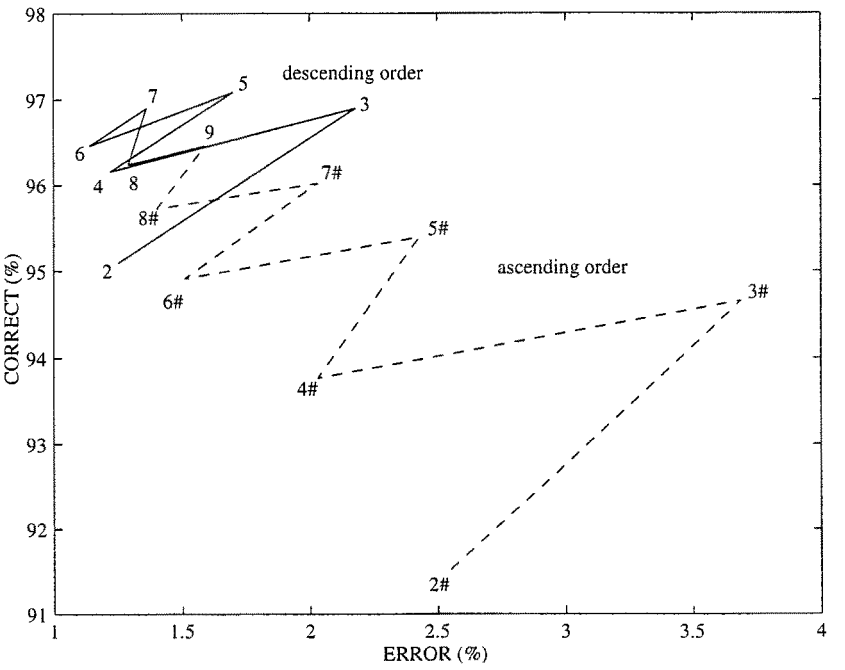

Fig. 1. Results from combinations of 9 classifiers on CEDAR database

2.3 Variations on Majority Vote

In the previous section, it has been shown that the performance of the combined decision is an increasing function of the number of experts, provided each expert can perform at an appropriate high level. The number of experts that can be used would naturally depend on practical limitations, and adding new experts with sufficiently high performance cannot always be readily accomplished. For this reason, we considered means to combine the existing experts in more optimal ways, and derived conditions as to which of the strategies would be preferable for a given objective.

Suppose an odd number of experts are available and a higher reliability is desired for the combination. This can be easily accomplished in one of two ways: to eliminate one of the experts from voting, or to double the vote of one of the experts by assigning a double weight to this vote. Either action would change the number of votes from an odd to an even number, resulting in more reliable results (at the cost of higher rejections). Of course, doubling one vote is equivalent to the addition of a dependent vote, but the results on the addition of one vote does not rely on independence. Analogously, when an even number of experts are available, the same options can be used to obtain an odd number of votes which would result in a higher recognition rate.

Intuitively, one would double the "best" and eliminate the "worst" algorithm, where these attributes are measured according to the correct and error rates. For algorithms with no rejections, the choice is obvious; otherwise the choice would depend on the trade-off between hgher recognition and lower error rates. Apart from this consideration, it remains to be resolved as to which alternative is better – to eliminate a vote or to double one.

Suppose the votes of n independent experts $v_1, v_2, \cdots, v_n$ are given, each with odds ratio $r_i of being correct, where 1 \leq i \leq n$. Then it has been shown [8] that the advantage of eliminating v_i versus doubling v_j depends on the pairwise products of the odds ratios. If

$$r_i r_j > r_k r_l \text{ for all } k, l \neq i, j,$$

then doubling v_j produces a higher recognition rate than eliminating v_i, for both even and odd values of n. When n is even and the reverse inequalities hold, then the opposite results are true. This result can be intuitively interpreted as follows.

It is logical to consider these alternatives only when r_i is relatively small and r_j is large. In addition, when v_j is doubled, then larger values of r_j would result in greater improvement. On the other hand, the elimination of v_i should lead to better results when r_i is smaller. Therefore, in the consideration of doubling v_j versus eliminating v_i, it is reasonable that the significant entity should be the product $r_i r_j$. When this product is large enough (as stipulated in the above condition), then r_j must be large and r_i not too small, when one may expect the results to be better when v_j is doubled.

In order to test the applicability of the theoretical results to a practical situation where the independence of experts cannot be completely assumed, we

consider the combinations of four out of six experts on the CENPARMI database from [2]. The performances of the experts are given in Table 3.

Table 3. Performances of experts on CENPARMI database

Expert	Recognition	Error	Rejection
E1	86.05	2.25	11.70
E2	92.85	2.45	4.70
E3	92.95	2.15	4.90
E4	93.90	1.60	4.50
E5	96.95	3.05	0.00
E6	98.30	1.70	0.00

The choice of four experts ensures that the above inequality or its reverse inequality will always be satisfied. Since experts E5 and E6 are highly correlated, combinations containing both of these experts are not considered. For each of the remaining nine combinations, v_1 refers to the first expert in the combination and v_4 the last, and the value of $d = r_1 r_4 - r_2 r_3$ is shown in Table 4 together with the recognition rates when v_4 is doubled and when v_1 is eliminated.

Table 4. Results of doubling v_4 versus eliminating v_1

Combination	$d = r_1 r_4 - r_2 r_3$	Recognition rate	
		Doubling v_4	Eliminating v_1
1 E1+2+3+4	-76.26	96.35	96.40
2 E1+2+3+5	24.86	97.20	97.10
3 E1+2+3+6	185.47	97.70	97.45
4 E1+2+4+5	-3.82	97.60	97.60
5 E1+2+4+6	156.78	98.00	97.70
6 E1+3+4+5	-6.88	97.70	97.40
7 E1+3+4+6	153.73	98.10	97.65
8 E2+3+4+5	209.93	97.85	97.40
9 E2+3+4+6	547.94	98.20	97.65

It is encouraging that the experimental results generally agree with the theoretical conclusion: when $d > 0$, doubling v_4 produces higher recognition rate than eliminating v_1, and vice versa. The exceptions are in combinations 4 and 6, in which d has very small magnitudes.

3 Application to Correlated Classifiers

Given that our theory is based on independence of classifiers, we wished to test its applicability when independence cannot be assumed. This need is apparent

when we consider that, since all classifiers consider the same scanned and digitized image to begin with, their decisions may not be completely independent. Therefore we experimented on correlated classifiers designed and implemented for automatic classification of human chromosomes. These classifiers make use of the same extracted features, and they differ only in the classification phase, which means they can be expected to be more highly correlated than classifiers which are developed separately using different features.

Chromosome classification is a 24-class problem in which the classes are 1-22, X and Y. Many different techniques have been developed for its automatic classification, and some common databases have been used for testing. Among them are the Copenhagen, Edinburgh, and Philadelphia datasets. Of these databases, the Copenhagen set containing 8106 chromosome images obtained from 180 peripheral blood cells, is the largest and contains the "cleanest" images.

For each chromosome, a 30-dimensional feature vector is extracted by Piper [13] and made available to other researchers. Our experiment uses the recognition results produced by 7 classifiers ([5], [6], [12], [11], [4], [16]). Since all the classifiers are based on the same features, they can be expected to be correlated, which was indicated in our experimental measurements.

The classification results of these seven algorithms are combined by majority vote, and Fig. 2 shows the combined results obtained.

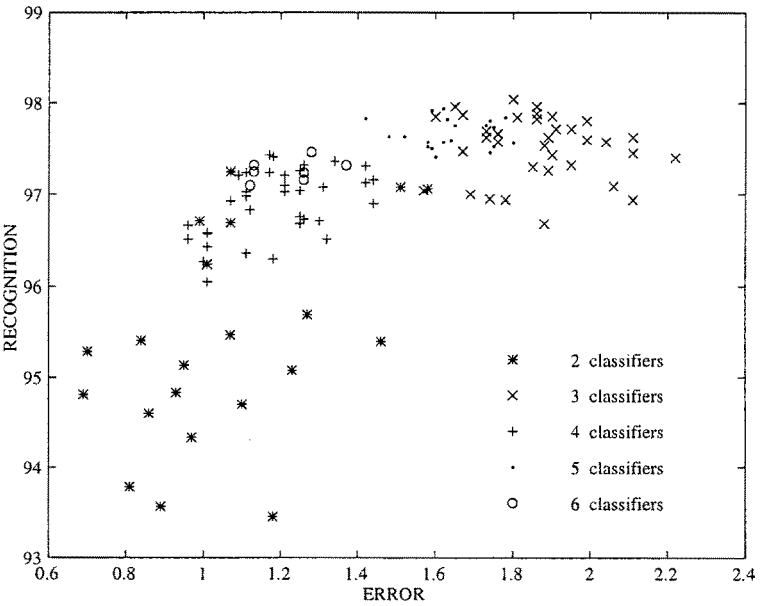

Fig. 2. Results from combinations of 7 classifiers on Copenhagen database

From this figure, it can be seen that despite their correlation, their majority vote results do show the tendencies derived theoretically based on the assumption of independence. For example, when one vote is added to an even number, the result moves towards the upper right (representing higher correct as well as error rates), while the movement is in the opposite direction when one more vote is added. At the same time, it is clear that the increase from 2 to 4, then to 6 experts results in mainly an upward trend (increase in recognition rate). This is in marked contrast to the leftward movement (decrease in error rate) produced by adding 2 experts to 3. The fact that these results are observed even in the absence of independence indicates that this assumption may be stronger than strictly needed for the theoretical conclusions.

Another result proved in [8] with the assumption of independence is that when the number n of experts is odd, then doubling a vote would tend to produce a higher recognition rate than eliminating a vote, even if the inequality is not satisfied for the odds ratios. However, this higher recognition rate would be accompanied by a higher error rate. These results are illustrated in Figure 3 and Figure 4.

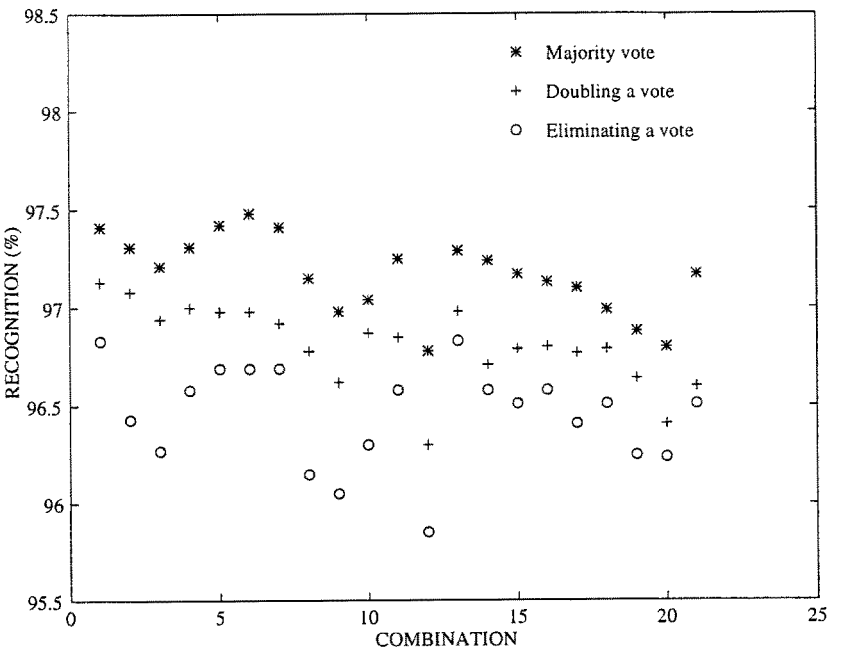

Fig. 3. Recognition rates from combinations of 5 classifiers on Copenhagen database

Figure 3 and Figure 4 represent the results obtained from all combinations of 5 out of the 7 chromosome classifiers. There are 21 such combinations, and we show the results of each combination, together with the results obtained from

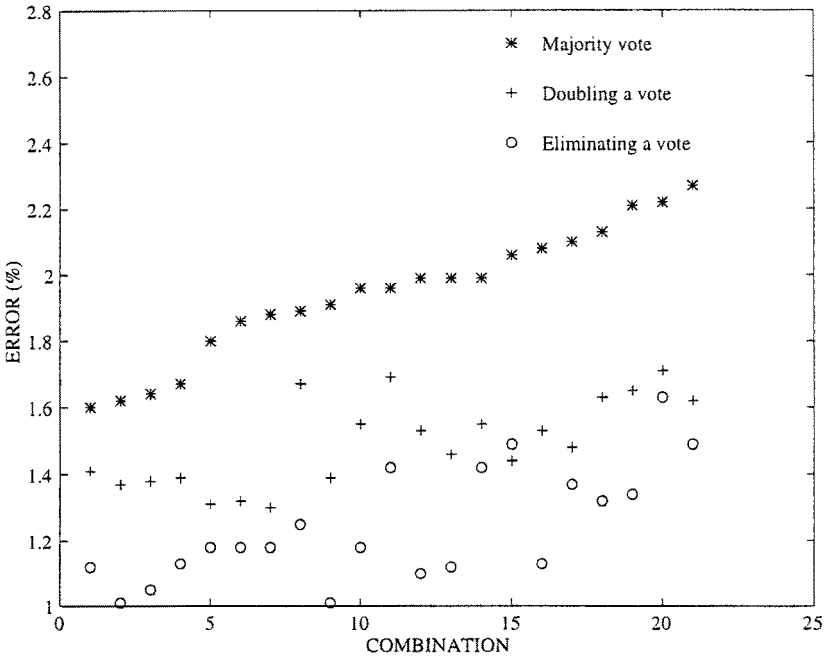

Fig. 4. Error rates from combinations of 5 classifiers on Copenhagen database

doubling the best and eliminating the weakest classifiers. The combinations are presented in order of performance of the original combination. It can be seen that the experimental results are in exact correspondence with the theoretical ones. For example, the doubling or elimination of one vote (to an even number of votes) produces lower recognition and error rates, while the elimination of the weakest vote results again in lower rates than the doubling of the best vote.

4 Conclusions

In this paper, we describe some theoretical results that we established on majority vote and show that these results are observable in actual experiments in pattern recognition even in the absence of independent decisions. These results enable the user to select an optimal strategy for combination of classifier decisions based on the desired objective.

Acknowledgements

This research was supported by the Natural Sciences and Engineering Research Council of Canada, the National Networks of Centres of Excellence program of Canada, and the FCAR program of the Ministry of Education of the province

of Québec. The first author is also supported by a research grant from the Hong Kong Institute of Education.

References

1. R. Fenrich and J. J. Hull. Concerns in creation of image databases. *Pre-Proceedings of the Third International Workshop in Frontiers on Handwriting Recognition*, Buffalo, NY, USA, May 1993, 112 - 121.
2. J. Franke, L. Lam, R. Legault, C. Nadal, and C. Y. Suen. Experiments with the cenparmi data base combining different classification approaches. *Pre-Proceedings of the Third International Workshop in Frontiers on Handwriting Recognition*, Buffalo, NY, USA, May 1993, 305 - 311.
3. T. K. Ho, J. J. Hull, and S. N. Srihari. Decision combination in multiple classifier systems. *IEEE Trans. PAMI,* **16** *(1994) 66 - 75.*
4. W.P. Sweeney Jr., M.T. Musavi, and J.N. Guidi. Classification of chromosomes using a probabilistic neural network. *Cytometry*, **16** (1994), 17–24.
5. P. Kleinschmidt, I. Mitterreiter, and J. Piper. Improved chromosome classification using monotonic functions of mahalanobis distance and the transportation method. *Mathematical Methods of Operations Research*, **40** (1994), 305–323.
6. P. Kleinschmidt, I. Mitterreiter, and C. Rank. A hybrid method for automatic chromosome karyotyping. *Pattern Recognition Letters*, **15** (1994), 87–96.
7. L. Lam, Y.-S. Huang, and C. Y. Suen. Combination of multiple classifier decisions for optical character recognition. *Handbook on Character Recognition and Digital Image Analysis*, eds. H. Bunke and P.S.P. Wang, World Scientific, 1997, 79 - 101.
8. L. Lam and C. Y. Suen. Application of majority voting to pattern recognition: an analysis of its behavior and performance. *IEEE Trans. Systems, Man, and Cybernetics*, **27** (1997) 553–568.
9. D.-S. Lee. A theory of classifier combination: the neural network approach. Doctoral Dissertation, State University of New York at Buffalo, 1995.
10. D.S. Lee and S.N. Srihari. Handprinted digit recognition: A comparison of algorithms, *pre-proc. 3rd int. workshop on frontiers in handwriting recognition*, buffalo, usa, may 1993, 153–162.
11. J. Piper. The effects of zero feature correlation assumption on maximum likelihood based classification of chromosomes. *Signal Processing*, **12** (1987), 49–57.
12. J. Piper. Classification of chromosomes constrained by expected class size. *Pattern Recognition Letters*, **4** (1986), 391–395.
13. J. Piper and E. Granum. On fully automatic feature measurement for banded chromosome classification. *Cytometry*, **10** (1989), 242–255.
14. T. Breuer R. Lindwurm and K. Kreuzer. Multi-expert system for handprint recognition. *Fifth Int. Workshop on Frontiers in Handwriting Recognition*, Colchester, UK, Sept. 1996, 125–129.
15. M. Revow, C. K. I. Williams, and G. E. Hinton. Using mixtures of deformable models to capture varaitions in hand printed digits. *Pre-Proceedings of the Third International Workshop in Frontiers on Handwriting Recognition*, Buffalo, NY, USA, May 1993, 142 - 152.
16. M. Tso, P. Kleinschmidt, I. Mitterreiter, and J. Graham. An efficient transportation algorithm for automatic chromosome karyotyping. *Pattern Recognition Letters*, **12** (1991), 117–126.
17. S. Yamaguchi, K. Nagata, T. Tsutsumida. F. Kimura, and A. Iwata. Study on multi-expert system for handprinted numeral recognition. *Fifth Int. Workshop on Frontiers in Handwriting Recognition*, Colchester, UK, Sept. 1996, 119–124.