Combining Classifiers

J Kittler, M Hatef* and R P W Duin+
Department of Electronic and Electrical Engineering
University of Surrey
Guildford
Surrey GU2 5XH, United Kingdom
and
* ERA Technology Ltd
Cleeve Road, Leatherhead KT22 7SA, United Kingdom
and
+ Faculty of Applied Physics
Delft University of Technology, Netherland

Abstract

We develop a common theoretical framework for combining classifiers which use distinct pattern representations and show that many existing schemes can be considered as special cases of compound classification where all the pattern representations are used jointly to make a decision. An experimental comparison of various classifier combination schemes demonstrates that the combination rule developed under the most restrictive assumptions - the sum rule - and its derivatives consistently outperform other classifier combinations schemes. A sensitivity analysis of the various schemes to estimation errors is carried out to show that this finding can be justified theoretically.

1. Introduction

The ultimate goal of designing pattern recognition systems is to achieve the best possible classification performance for the task at hand. This objective traditionally led to the development of different classification schemes for any pattern recognition problem to be solved. The results of an experimental assessment of the different designs would then be the basis for choosing one of the classifiers as a final solution to the problem. It had been observed in such design studies, that although one of the designs would yield the best performance, the sets of patterns misclassified by the different classifiers would not necessarily overlap. This suggested that different classifier designs potentially offered complementary information about the patterns to be classified which could be harnessed to improve the performance of the selected classifier.

These observations motivated the relatively recent interest in combining classifiers[1]–[13]. The idea is not to rely on a single decision making scheme. Instead, all the designs, or their subset, are used for decision making by combining their individual opinions to derive a consensus decision. Various classifier combination schemes have been devised and it has been experimentally demonstrated that some of them consistently outperform a single best classifier. However, there is presently inadequate understanding why some combination schemes are better than others and in what circumstances.

From the point of view of their analysis, there are basically two classifier combination scenarios. In the first scenario, all the classifiers use the same representation of the input pattern. A typical example of this category is a set of k-nearest neighbour classifiers, each using the same measurement vector, but different classifier parameters (number of nearest neighbours k, or distance metrics used for determining the nearest neighbours).

In the second scenario each classifier uses its own representation of the input pattern. In other words, the measurements extracted from the pattern are unique to each classifier. An important application of combining classifiers in this scenario is the possibility to integrate physically different types of measurements / features. In this case it is no longer possible to consider the computed aposteriori probabilities to be estimates of the same functional value, as the classification systems operate in different measurement spaces.

We provide a theoretical underpinning of the existing classifier combination schemes. Furthermore, our analysis of the sensitivity of these schemes to estimation errors en-

1015-4651/96 $5.00 © 1996 IEEE
Proceedings of ICPR '96

hances the understanding of their properties. As a byproduct, we also offer a methodological machinery which can be used for developing other classifier combination strategies and for predicting their behaviour.

The paper is organised as follows. In Section 2 we formulate the classifier combination problem and introduce the necessary notation. In this section we also derive the basic classifier combination schemes: the product rule and the sum rule. These two basic schemes are then developed into other classifier combination strategies in Section 3. The combination rules derived in Sections 2 and 3 are experimentally compared in Section 4. Section 5 investigates the sensitivity of the basic classifier combination rules to estimation errors. Finally, Section 6 summarises the main results of the paper and offers concluding remarks.

2. Theoretical Framework

Consider a pattern recognition problem where pattern Z is to be assigned to one of the m possible classes $\{\omega_1, ..., \omega_m\}$. Let us assume that we have R classifiers each representing the given pattern by a distinct measurement vector. Denote the measurement vector used by the i-th classifier by $\mathbf{x}_i$. In the measurement space each class ω_k is modelled by the probability density function $p(\mathbf{x}_i|\omega_k)$ and its a priori probability of occurrence is denoted $P(\omega_k)$. We shall consider the models to be mutually exclusive which means that only one model can be associated with each pattern.

Now according to the Bayesian theory, given measurements $\mathbf{x}_i, i = 1, ..., R$, the pattern, Z, should be assigned to class ω_j, i.e. its label θ should assume value $\theta = \omega_j$, provided the aposteriori probability of that interpretation is maximum, i.e.

$$assign \qquad \theta \rightarrow \omega_j \qquad if$$

$$P(\theta = \omega_j|\mathbf{x}_1, ..., \mathbf{x}_R) = \max_k P(\theta = \omega_k|\mathbf{x}_1, ..., \mathbf{x}_R) \quad (1)$$

The Bayesian decision rule (1) states that in order to utilise all the available information correctly to reach a decision, it is essential to compute the probabilities of the various hypotheses by considering all the measurements simultaneously. This is, of course, a correct statement of the classification problem but it may not be a practicable proposition. The computation of the aposteriori probability functions would depend on the knowledge of high order measurement statistics described in terms of joint probability density functions $p(\mathbf{x}_1, ..., \mathbf{x}_R|\theta = \omega_k)$ which would be difficult to infer. We shall therefore attempt to simplify the above rule and express it in terms of decision support computations performed by the individual classifiers, each exploiting only the information conveyed by vector $\mathbf{x}_i$. We shall see that this will not only make rule (1) computationally manageable, but also it will lead to combination rules which are commonly used in practice. Moreover, this approach will provide a scope for the development of a range of efficient classifier combination strategies.

Let us rewrite the aposteriori probability $P(\theta = \omega_k|\mathbf{x}_1, ..., \mathbf{x}_R)$ using the Bayes theorem. We have

$$P(\theta = \omega_k|\mathbf{x}_1, ..., \mathbf{x}_R) = \frac{p(\mathbf{x}_1, ..., \mathbf{x}_R|\theta = \omega_k)P(\omega_k)}{p(\mathbf{x}_1, ..., \mathbf{x}_R)} \quad (2)$$

where $p(\mathbf{x}_1, ..., \mathbf{x}_R)$ is the unconditional measurement joint probability density. The latter can be expressed in terms of the conditional measurement distributions as and therefore, in the following, we can concentrate only on the numerator terms of (2).

Product Rule

As already pointed out, $p(\mathbf{x}_1, ..., \mathbf{x}_R|\theta = \omega_k)$ represents the joint probability distribution of the measurements extracted by the classifiers. Since the representations used by the classifiers are distinct, it may be true for some applications that these measurements will be conditionally statistically independent. We will investigate the consequences and write

$$p(\mathbf{x}_1, ..., \mathbf{x}_R|\theta = \omega_k) = \Pi_{i=1}^R p(\mathbf{x}_i|\theta = \omega_k) \quad (3)$$

where $p(\mathbf{x}_i|\theta = \omega_k)$ is the measurement process model of the i-th representation. Substituting from (3) into (2) we find

$$P(\theta = \omega_k|\mathbf{x}_1, ..., \mathbf{x}_R) = \frac{P(\omega_k)\Pi_{i=1}^R p(\mathbf{x}_i|\theta = \omega_k)}{\sum_j^m P(\omega_j)\Pi_{i=1}^R p(\mathbf{x}_i|\theta = \omega_j)} \quad (4)$$

and using (4) in (1) we obtain the decision rule

$$assign \qquad \theta \rightarrow \omega_j \qquad if$$

$$P(\omega_j)\Pi_{i=1}^R p(\mathbf{x}_i|\theta = \omega_j) = \max_{k=1}^m P(\omega_k)\Pi_{i=1}^R p(\mathbf{x}_i|\theta = \omega_k) \quad (5)$$

or in terms of the aposteriori probabilities yielded by the respective classifiers

$$assign \qquad \theta \rightarrow \omega_j \qquad if$$

$$P^{-(R-1)}(\omega_j)\Pi_{i=1}^R P(\theta = \omega_j|\mathbf{x}_i) = \max_{k=1}^m P^{-(R-1)}(\omega_k)\Pi_{i=1}^R P(\theta = \omega_k|\mathbf{x}_i) \quad (6)$$

The decision rule (6) quantifies the likelihood of a hypothesis by combining the aposteriori probabilities generated by the individual classifiers by means of a product rule. It is effectively a severe rule of fusing the classifier outputs as it is sufficient for a single recognition engine to inhibit a particular interpretation by outputting a close to zero probability for it. As we shall see below, this has a rather undesirable implication on the decision rule combination as all the classifiers, in the worst case, will have to provide their respective opinions for a hypothesised class identity to be accepted or rejected.

Sum Rule

Let us consider decision rule (6) in more detail. In some applications it may be appropriate further to assume that the aposteriori probabilities computed by the respective classifiers will not deviate dramatically from the prior probabilities. This is a rather strong assumption but it may be readily satisfied when the available observational discriminatory information is highly ambiguous due to high levels of noise. In such a situation we can assume that the aposteriori probabilities can be expressed as

$$P(\theta = \omega_k|\mathbf{x}_i) = P(\theta = \omega_k)(1 + \delta_{ki}) \quad (7)$$

where δ_{ki} satisfies $\delta_{ki} << 1$. Substituting (7) for the aposteriori probabilities in (6) we find

$$P^{-(R-1)}(\omega_k)\Pi_{i=1}^{R} P(\theta = \omega_k|\mathbf{x}_i) = P(\omega_k)\Pi_{i=1}^{R}(1 + \delta_{ki}) \quad (8)$$

If we expand the product and neglect any terms of second and higher order we can approximate the right hand side of (8) as

$$P(\omega_k)\Pi_{i=1}^{R}(1 + \delta_{ki}) = P(\omega_k) + P(\omega_k)\sum_{i=1}^{R}\delta_{ki} \quad (9)$$

Substituting (9) and (7) into (6) we obtain a sum decision rule

$$assign \qquad \theta \rightarrow \omega_j \qquad if$$

$$(1 - R)P(\omega_j) + \sum_{i=1}^{R} P(\omega_j|\mathbf{x}_i) = \max_{k=1}^{m}[(1 - R)P(\omega_k) + \sum_{i=1}^{R} P(\omega_k|\mathbf{x}_i)] \quad (10)$$

In the following section, we shall develop specific classifier combination strategies based on decision rules (6) and (10).

3. Classifier Combination Strategies

Many commonly used classifier combination strategies can be developed from rules (6) and (10) by noting that

$$\Pi_{i=1}^{R} P(\theta = \omega_k|\mathbf{x}_i) \leq min_{i=1}^{R} P(\theta = \omega_k|\mathbf{x}_i) \leq \frac{1}{R}\sum_{i=1}^{R} P(\theta = \omega_k|\mathbf{x}_i) \leq max_{i=1}^{R} P(\theta = \omega_k|\mathbf{x}_i) \quad (11)$$

The relationship (11) suggests that the product and sum combination rules can be approximated by the above upper or lower bounds, as appropriate. Furthermore, the hardening of the aposteriori probabilities $P(\theta = \omega_k|\mathbf{x}_i)$ to produce binary valued functions Δ_{ki} as

$$\Delta_{ki} = \begin{cases} 1 & \text{if } P(\theta = \omega_k|\mathbf{x}_i) = max_{j=1}^{m} P(\theta = \omega_j|\mathbf{x}_i) \\ 0 & \text{otherwise} \end{cases} \quad (12)$$

results in combining decision outcomes rather than combining aposteriori probabilities. These approximations lead to the following rules(for details see [16]:

$$assign \qquad \theta \rightarrow \omega_j \qquad if$$

- Max Rule

$$max_{i=1}^{R} P(\theta = \omega_j|\mathbf{x}_i) = \max_{k=1}^{m} max_{i=1}^{R} P(\theta = \omega_k|\mathbf{x}_i) \quad (13)$$

- Min Rule

$$min_{i=1}^{R} P(\theta = \omega_j|\mathbf{x}_i) = \max_{k=1}^{m} min_{i=1}^{R} P(\theta = \omega_k|\mathbf{x}_i) \quad (14)$$

- Median Rule

$$med_{i=1}^{R} P(\omega_j|\mathbf{x}_i) = \max_{k=1}^{m} med_{i=1}^{R} P(\omega_k|\mathbf{x}_i) \quad (15)$$

- Majority Vote Rule

$$\sum_{i=1}^{R}\Delta_{ji} = \max_{k=1}^{m}\sum_{i=1}^{R}\Delta_{ki} \quad (16)$$

4. Experimental Results

The problem of handwritten character recognition is used to investigate the different schemes of combining classifiers. We used 18468 samples as training set and 2213 samples as test set. The data base is the CEDAR-CDROM which are images scanned from dead letter envelopes provided by the US postal service.

Four different classifiers have been used: structural[14], gausian , neural network and hidden markov model [15].

Six different combination schemes are applied and their results are compared. These schemes can be divided into two groups according to the format of the individual classifiers used by the combiner. Hard-level combination uses the output of the classifier after it is hardened(binarised). Soft-level on the other hand uses the estimates of the a posteriori probability of the class by each classifier. The majority vote combiner is a representative of the first category while the five different operators are the soft-level combiners. Table 1 shows the results of classification of the individual classifiers as well as the different combining schemes.

Note that the worst results are achieved when using the *product* rule which has exactly the same level of performance as the *min* rule. The results using these two rules are worse than any of the individual classifiers and the reason is that if any of the classifiers reports the correct class aposteriori probability as 0 the output will be 0 and the correct class cannot be identified. Therefore the final result reported by the combiner in such cases is either the wrong class(worst

Classifier	Classification rate %
Structural:	90.85
Gaussian:	93.93
Neural Net:	93.20
HMM:	94.77
Majority Vote:	97.96
Sum rule:	98.05
Max rule:	93.93
Min rule:	86.00
Product rule:	84.69
Median rule:	98.19

Table 1. **The classification rate of individulal classifier and different combiners.**

case) or a reject when all of the classes are assigned zero aposteriori probability. Another interesting outcome from our experiments is that the *Mean* rule as well as the *median* rule have the best classification results. *Majority vote* rule is very close in performance to the mean and median rules.

5. Error Sensitivity

A somewhat surprising outcome of the experimental comparison of the classifier combination rules reported in Section 4 is that the sum rule (10), which has been developed under the strongest assumptions, namely those of

- conditional independence of the respective representations used by the individual classifiers, and
- classes being highly ambiguous (observations enhance the apriori class probabilities only slightly)

appears to produce the most reliable decisions. In this section we shall investigate the sensitivity of the two generic classifier combination rules - the product rule (6) and the sum rule (10) to estimation errors. We shall show that the sum rule is much less affected by estimation errors. This theoretically established behaviour is consistent with the experimental findings.

In the developments in Sections 2 and 3 we assumed that the aposteriori class probabilities $P(\omega_j|\mathbf{x}_i)$, in terms of which the various classifier combination rules are defined, are computed correctly. In fact each classifier i will produce only an estimate of this probability, which we shall denote $\hat{P}(\omega_j|\mathbf{x}_i)$. The estimate deviates from the true probability by error e_{ji}, i.e.

$$\hat{P}(\omega_j|\mathbf{x}_i) = P(\omega_j|\mathbf{x}_i) + e_{ji} \tag{17}$$

It is these estimated probabilities that enter the classifier combination rules rather than the true probabilities.

Let us now consider the effect of the estimation errors on the classifier combination rules. Substituting (17) into (6) we have

$$assign \qquad \theta \rightarrow \omega_j \qquad if$$

$$P^{-(R-1)}(\omega_j)\Pi_{i=1}^{R}[P(\theta=\omega_j|\mathbf{x}_i)+e_{ji}] = \max_{k=1}^{m} P^{-(R-1)}(\omega_k)\Pi_{i=1}^{R}[P(\theta=\omega_k|\mathbf{x}_i)+e_{ki}] \tag{18}$$

Under the assumption that $e_{ki} << P(\theta=\omega_k|\mathbf{x}_i)$ which is rather strong and may not represent the worst case scenario, and further assuming that $P(\theta=\omega_k|\mathbf{x}_i) \neq 0$ we can rearrange the product term as

$$\Pi_{i=1}^{R}[P(\theta=\omega_k|\mathbf{x}_i)+e_{ki}] = [\Pi_{i=1}^{R}P(\theta=\omega_k|\mathbf{x}_i)]* \\ * \Pi_{i=1}^{R}[1+\frac{e_{ki}}{P(\theta=\omega_k|\mathbf{x}_i)}] \tag{19}$$

which can then be linearised as

$$\Pi_{i=1}^{R}[P(\theta=\omega_k|\mathbf{x}_i)+e_{ki}] = \\ [\Pi_{i=1}^{R}P(\theta=\omega_k|\mathbf{x}_i)][1+\sum_{i=1}^{R}\frac{e_{ki}}{P(\theta=\omega_k|\mathbf{x}_i)}] \tag{20}$$

Substituting (20) into (18) we get

$$assign \qquad \theta \rightarrow \omega_j \qquad if$$

$$[P^{-(R-1)}(\omega_j)\Pi_{i=1}^{R}P(\theta=\omega_j|\mathbf{x}_i)][1+\sum_{i=1}^{R}\frac{e_{ji}}{P(\theta=\omega_j|\mathbf{x}_i)}] = \\ = \max_{k=1}^{m}[P^{-(R-1)}(\omega_k)\Pi_{i=1}^{R}P(\theta=\omega_k|\mathbf{x}_i)]... \\ [1+\sum_{i=1}^{R}\frac{e_{ki}}{P(\theta=\omega_k|\mathbf{x}_i)}] \tag{21}$$

Comparing (6) and (21) it is apparent that each term (class ω_k hypothesis) in the *error free* classifier combination rule (6) is affected by error factor

$$[1+\sum_{i=1}^{R}\frac{e_{ki}}{P(\theta=\omega_k|\mathbf{x}_i)}] \tag{22}$$

A similar analysis of the sum rule (10) results in the corresponding error factor given as

$$[1+\frac{\sum_{i=1}^{R}e_{ki}}{\sum_{i=1}^{R}[P(\omega_k|\mathbf{x}_i)}] \tag{23}$$

Comparing error factors (22) and (23) it transpires that the sensitivity to errors of the former is much more dramatic than that of the latter. Note that since the aposteriori class probabilities are less than unity, each error e_{ki} in (22) is amplified by $\frac{1}{P(\omega_k|\mathbf{x}_i)}$. The compounded effect of all these amplified errors is equivalent to their sum. In contrast in the sum rule the errors are not amplified. On the contrary, their

compounded effect, which is also computed as a sum, is scaled by the the sum of the aposteriori probabilities. For the most probable class this sum is likely to be greater than one which will result in the dampening of the errors. Thus the sum decision rule is much more resilient to estimation errors as we observed experimentally in Section 3. It follows, therefore, that the sum classifier combination rule is not only a very simple and intuitive technique of improving the reliability of decision making based on different classifier opinions but it is also remarkably robust. Moreover, as the sum rule is a basis of a number of other classifier combination strategies, this error resilience extends also to these rules. Specifically, the *max rule, majority vote rule*, and the *median rule* should inherit this robustness and this has been confirmed by the experimental results.

6. Conclusions

The problem of combining classifiers which use different representations of the patterns to be classified was studied. We have developed a common theoretical framework for classifier combination and showed that many existing schemes can be considered as special cases of compound classification where all the pattern representations are used jointly to make a decision. We have demonstrated that under different assumptions and using different approximations we can derive the commonly used classifier combination schemes such as the product rule, sum rule, min rule, max rule, median rule and majority voting. The various classifier combination schemes were compared experimentally. A surprising outcome of the comparative study was that the combination rule developed under the most restrictive assumptions - the sum rule - and its derivatives consistently outperformed other classifier combinations schemes. To explain this empirical finding, we investigated the sensitivity of the various schemes to estimation errors. The sensitivity analysis has shown that the sum rule is most resilient to estimation errors.

7. Acknowledgements

This work was supported by the Science and Engineering Research Council, UK (GR/K68165). The authors would like to thank Andrew Elms for making available the classification results obtained using his HMM character recogniser.

References

[1] J. Cao, M. Ahmadi, and M. Shridhar, Recognition of handwritten numerals with multiple feature and multistage classifier, *Pattern Recognition*, vol. 28, no. 2, 1995, 153-160.

[2] S.B. Cho and J.H. Kim, "Multiple network fusion using fuzzy logic," *IEEE Transactions on Neural Networks*, vol. 6, no. 2, pp. 497-501, 1995.

[3] J. Franke and E. Mandler, A Comparison of Two Approaches for Combining the Votes of Cooperating Classifiers. *Proceedings 11th IAPR International Conference on Pattern Recognition, Volume II, Conference B: Pattern Recognition Methodology and Systems*, 1992, 611-614.

[4] L.K. Hansen and P. Salamon, "Neural network ensembles", *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 12, no. 10, pp. 993- 1001, 1990.

[5] Hashem and B. Schmeiser, Improving model accuracy using optimal linear combinations of trained neural networks, *IEEE Transactions on Neural Networks*, vol. 6, no. 3, 1995, 792-794.

[6] T.K. Ho, J.J. Hull, and S.N. Srihari, "Decision combination in multiple classifier systems",*IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 16, no. 1, pp. 66-75, 1994.

[7] F. Kimura and M. Shridhar, Handwritten numerical recognition based on multiple algorithms, *Pattern Recognition*, vol. 24, no. 10, 1991, 969-983.

[8] A. Krogh and J. Vedelsby, "Neural network ensembeles, cross validation, and active learning," in:*Advances in neural information processing systems 7, ed. G. Tesauro, D.S. Touretzky, T.K. Leen*, MIT Press, Cambridge MA, 1995.

[9] G. Rogova, Combining the results of several neural network classifiers,*Neural Networks*, vol. 7, no. 5, 1994, 777-781.

[10] V. Tresp, M. Taniguchi, Combining estimators using non-constant weighting functions, in: *G. Tesauro, D.S. Touretzky, T.K. Leen, eds., Advances in neural information processing systems 7*, MIT Press, Cambridge MA, 1995

[11] C.H. Tung, H.J. Lee, and J.Y. Tsai, Multi-stage pre-candidate selection in handwritten Chinese character recognition systems, *Pattern Recognition*, vol. 27, no. 8, 1994, 1093-1102.

[12] D.H. Wolpert, Stacked generalization, *Neural Networks*, vol. 5, no. 2, 1992, 241-260.

[13] L. Xu, A. Krzyzak, C.Y. Suen, Methods of combining multiple classifiers and their applications to handwriting recognition, *IEEE Trans.* SMC, vol. 22, no. 3, 1992, pp. 418-435.

[14] M. Hatef and J. Kittler, Constraining probabilistic relaxation with symbolic attributes, *Proc. 6th Internat Conference on Computer Analysis of Images and Patterns*, V Hlavac and R Sara (Eds), 862-867, Prague 1995.

[15] A. J. Elms, A connected character recogniser using Level Building of HMMs, *Proceedings 12th IAPR International Conference and Neural Networks*, Volume II, Conference B: Pattern Recognition Methodology and Systems 1994, 439-441.

[16] M. Hatef and J. Kittler and R. P. Duin, Combining Multiple clssifiers, *VSSP Technichal Report*, University of Surrey, 1996.