Combined Classifier Optimisation via Feature Selection

David Windridge and Josef Kittler

Centre for Vision, Speech and Signal Processing
Dept. of Electronic & Electrical Engineering, University of Surrey
Guildford, GU2 5XH Surrey, United Kingdom
Telephone: +44 1483 876043
d.windridge@eim.surrey.ac.uk

Abstract. We propose a new method for performance-constraining the feature selection process as it relates to combined classifiers, and assert that the resulting technique provides an alternative to the more familiar optimisation methodology of weight adjustment. The procedure then broadly involves the prior selection of features via performance-constrained sequential forward selection applied to the classifiers individually, with a subsequent forward selection process applied to the classifiers acting in combination, the selection criterion in the latter case deriving from the combined classification performance. We also provide a number of parallel investigations to indicate the performance enhancement expected of the technique, including an exhaustive weight optimisation procedure of the customary type, as well as an alternative backward selection technique applied to the individually optimised feature sets.

1 Introduction

The non-overlapping of the misclassification errors of very distinct methods of classification has lead to the realisation that, in general, no one method of classification can circumscribe all aspects of a typical real-world classification problem, prompting, in consequence, the investigation of a variety of combinatorial methods in a bid to improve classification performance [eg 1-6]. Historically, such methods have in common that they operate at the level of the compound classifiers' output, typically combining the disparate PDFs in some fashion (eg the majority vote and weighted mean techniques familiar to the pattern-recognition and sensor-fusion communities), and, as such, not having any direct influence on the compositional character of the feature set presented to each of the classifiers. We seek to address this deficit by attempting to obtain a near optimal feature set for the *combined* classifiers, in distinction to the optimal set for the *individual* classifiers, treating the latter as a starting point for this objective. Hence, by implication, this paper may also be considered an investigation into classifier combination as a constraint on feature selection, this being an issue in its own right; our primary aim, however, will be to improve combined classifier performance.

F.J. Ferri et al. (Eds.): SSPR&SPR 2000, LNCS 1876, pp. 687–695, 2000.

2 Format of the Investigation

We opt for the most straightforward method of classifier combination; that of obtaining the mean of the estimated posterior probability distributions of each of the classifiers composing the combination in order to elaborate our technique [cf eg 7]. It is not expected that this choice will substantially impact on the broad pattern of results that we find in section 4; in particular, the predominance of the sequential forward search method over the backward search; further results or mathematical argument would be required to provide absolute proof of this, however.

The criterion function for feature selection appropriate to the outlined combinatorial method is then simply the inverse of the misclassification rate arising from this mean of the estimated posterior probabilities in relation to the given feature set. The technique of feature set modification common to all of the various strands of the investigation is then the sequential selection of each of the features in turn from a bank of permissible features appropriate both to the particular classifier under consideration, as well as the method of feature set selection (ie unchosen features in the case of forward selection and unremoved features in the case of backward selection), respectively adding or subtracting the chosen feature from the existing set presented to that classifier. The previously specified criterion function is then calculated for the combination, with the remaining classifiers maintaining their existing feature set (or when individual classifiers are to be considered in isolation, as below, the criterion function will instead derive from the selected classifier alone). The feature/classifier combination with the most advantageous criterion function is then appended/removed, as appropriate to the method of selection, from the list of permissible permutations. Thus feature repetition between (but not amongst) classifiers is an inherent possibility in all but the case of the classifiers considered on an individual basis. There is then maximal freedom in the allocation of features, given that all of the various processes constituting the investigation are, in the broadest sense, sequential selection methods and thus subject to the "nesting effect" [cf 8] wherein features, once selected, lack any mechanism for removal from the set presented to the classifier (with an equivalent, though inverted, problem for backward selection). This situation is slightly mitigated in the particular case of sequential backward selection applied to the forwardly pre-optimised feature sets outlined below, although results in section 4 indicate that this is not the favoured amongst the various possibilities in performance terms. Thus, all of the following techniques (the most effective one of which, by default, constituting the proposed method of combined classifier optimisation, with the remainder to be considered only as relative performance indicators) are invariably sub-optimal; any such predominating method may therefore be most usefully treated as a relatively computational inexpensive addition to the repertoire of techniques for combined-classier optimisation, lying somewhere between exhaustive classifier weighting optimisation and exhaustive feature permutation optimisation, in terms of both execution time and performance.

The various methods of feature selection optimisation constituting the investigation are therefore:

1. Sequential forward selection employing a combined classifier criterion function and permitting unrestricted repetition of features between classifiers.
2. Sequential backward selection employing a combined classifier criterion function, commencing with complete feature sets for all of the constitutive classifiers.
3. Sequential forward selection applied to each of the classifiers individually, employing the inverse of the misclassification rate of the estimated posterior probability distribution as the criterion function in each case.
4. Sequential forward selection (employing the combined classifier criterion function) applied to the individually optimised feature sets for all constituent classifiers of the combination as derived from investigation number 3.
5. Sequential *backward* selection (employing the combined classifier criterion function) applied to the individually optimised feature sets for all classifiers in the combination, as derived from investigation number 3.
6. As a relative measure of the classification performance improvement attributable to the above processes, we supply an additional exhaustive weight optimisation to be applied to the individually optimised classifier/feature combinations derived from investigation number 3, acting in combination via the usual mechanism (mean, and hence thus now *weighted* mean). In this scenario the feature sets are not subject to change after their initial derivation by independent sequential forward selection, the weight modification being the sole source of performance optimisation, the exhaustivity of which being guaranteed by a series of nested loops, within which every permutation of PDF weight values (to within a specified resolution parameter) is inherently tested. The efficiency of this method (despite the series of nested loops) derives from the fact that the estimated posterior probabilities for each of the classifiers need not be re-derived for every iteration of the loop, the weights simply acting in multiple combination with the estimated class PDFs. This is not the case for the previously listed techniques, all of which have therefore a substantially greater (if not necessarily prohibitive) execution time.

3 Nature of Implementation

3.1 The Data

The data employed throughout the investigation consists in a twined set of expertly-classified geological survey data, one real and the other simulated, the latter simulation occurring at a stage *prior* to the application of the various pattern recognition methods from which the features are derived, and thus providing a measure of the distinction between conceptual and by-sight classification. In regard to our investigation, however, the essential difference between the two

data sets may be considered simply in terms of their class separability; the simulated data set exhibits this quality to a far greater degree than the real data, for which the class membership ambiguity is considerable.

The nature of the image processing on the two data sets (which parallelled each other exactly) consisted in a battery of 26 cell-based processes for texture characterisation, chosen without regard to the particular nature of the classification problem. Thus, from the outset, a particularly high feature redundancy was anticipated for the corresponding 26-dimensional pattern vector.

3.2 The Classifiers

Four classifiers constituted the combination, chosen to collectively represent the gamut of classification philosophies. They are:

1. **Nearest Neighbour Classifier:**
   This is a standard "1-NN" nearest neighbour classifier with Euclidean metric, adopted in place of the more reliable k-NN set of classifiers for reasons of efficiency, as well as conformity with the objective of bringing about approximate parity of misclassification rates amongst the various classifiers.
2. **Neural Net Classifier:**
   A Bayesian neural net classifier consisting of 3 hidden layers.
3. **Normal PDF Classifier:**
   A Bayesian classifier employing a normal probability density function estimator.
4. **Quadratic PDF Classifier:**
   As above, but employing a quadratic polynomial fitting function for the density estimation.

4 Results

The results of the six investigations are tabulated below for the real and synthetic data sets, respectively, with the training and test set data in both cases comprised of 1000 (of a possible 10000) random samples of their respective originals. The processing stages are in each case listed up to the point immediately preceding the termination of the procedure, at the point at which the peak of performance has been determined as being such (which is to say, exactly one iteration after the peak itself is reached). This approach has been adopted primarily as an efficiency measure, there being no reason to suppose that there might exist further modalities to the performance distribution beyond this single peak; test procedures without this imposed terminating condition have tended to confirm the validity of the supposition.

5 Conclusions

A consideration of the experimental findings set out in section 4 would appear to suggest that method number 4 constitutes the most effective of the tested

Table 1. Results of investigation number 1. (Real Data)

Order of feature addition:	1st	2nd	3rd	4th	5th	6th
Feature added:	22	24	16	9	3	1
Classifier to which feature added:	4	3	3	3	3	2
Probability of misclassification:	0.0622269	0.0442208	0.0386601	0.0264795	0.0169469	0.0112538

Table 2. Results of investigation number 1. (Synthetic Data)

Order of feature addition:	1st	2nd	3rd	4th	5th
Feature added:	21	10	1	7	19
Classifier to which feature added:	4	3	3	3	3
Probability of misclassification:	0.269688	0.184305	0.165239	0.144101	0.134706

Table 3. Results of investigation number 2. (Real Data)

Order of feature removal:	1st	2nd	3rd	4th	5th	6th	7th	8th
Feature removed:	21	13	2	7	12	17	20	21
Classifier from which feature removed:	1	2	2	4	4	4	4	3
Probability of misclassification:	0.071	0.062	0.052	0.049	0.047	0.046	0.043	0.042

Table 4. Results of investigation number 2. (Synthetic Data)

Order of feature removal:	1st	2nd	3rd	4th	5th	6th	7th	8th	9th	10th	11th
Feature removed:	16	17	13	25	11	7	14	4	15	10	14
Classifier from which feature removed:	1	1	2	4	4	4	3	4	4	3	1
Probability of misclassification:	0.35	0.33	0.28	0.28	0.27	0.27	0.26	0.26	0.26	0.25	0.24

Table 5. Results of investigation number 3. (Classifier 1, Real Data)

Order of feature addition:	1st	2nd	3rd
Feature added:	2	4	16
Probability of misclassification:	0.15848	0.0581226	0.0382629

Table 6. Results of investigation number 3. (Classifier 1, Synthetic Data)

Order of feature addition:	1st	2nd	3rd	4th
Feature added:	22	24	1	25
Probability of misclassification:	0.464907	0.201161	0.158607	0.118265

Table 7. Results of investigation number 3. (Classifier 2, Real Data)

Order of feature addition:	1st
Feature added:	1
Probability of misclassification:	0.102608

Table 8. Results of investigation number 3. (Classifier 2, Synthetic Data)

Order of feature addition:	1st	2nd	3rd	4th
Feature added:	22	24	23	20
Probability of misclassification:	0.30257	0.187621	0.159713	0.157917

Table 9. Results of investigation number 3. (Classifier 3, Real Data)

Order of feature addition:	1st
Feature added:	1
Probability of misclassification:	0.109361

Table 10. Results of investigation number 3. (Classifier 3, Synthetic Data)

Order of feature addition:	1st	2nd	3rd	4th	5th
Feature added:	1	17	12	14	21
Probability of misclassification:	0.472368	0.347748	0.28005	0.229898	0.212351

Table 11. Results of investigation number 3. (Classifier 4, Real Data)

Order of feature addition:	1st	2nd
Feature added:	22	15
Probability of misclassification:	0.0622269	0.0558718

Table 12. Results of investigation number 3. (Classifier 4, Synthetic Data)

Order of feature addition:	1st	2nd	3rd
Feature added:	21	4	22
Probability of misclassification:	0.269688	0.200055	0.153634

Table 13. Results of investigation number 4. (Real Data)

Order of feature addition:	Initial State	1st	2nd	3rd	4th	5th	6th
Feature added:	–	15	18	23	2	16	20
Classifier to which feature added:	–	2	1	3	3	3	3
Probability of misclassification:	0.076525	0.016152	0.010459	0.010194	0.010082	0.010078	0.010062

Table 14. Results of investigation number 4. (Synthetic Data)

Order of feature addition:	Initial State	1st	2nd	3rd
Feature added:	–	13	11	20
Classifier to which feature added:	–	4	3	3
Probability of misclassification:	0.160403	0.0965736	0.0893893	0.0862117

Table 15. Results of investigation number 5. (Real Data)

Order of feature removal:	Initial State	1st
Feature removed:	–	1
Classifier from which feature removed:	–	3
Probability of misclassification:	0.0765259	0.0271415

Table 16. Results of investigation number 5. (Synthetic Data)

Order of feature removal:	Initial State	1st	2nd
Feature removed:	–	21	25
Classifier from which feature removed:	–	3	1
Probability of misclassification:	0.160403	0.105554	0.103482

Table 17. Results of investigation number 6. (Real Data)

Classifier:	1	2	3	4
Final weight combination:	0.21	0.00	0.05	0.63
Unweighted performance:	0.0765259			
Final weighted performance:	0.0164173			

Table 18. Results of investigation number 6. (Synthetic Data)

Classifier:	1	2	3	4
Final weight combination:	0.57	0.00	0.22	0.95
Unweighted performance:	0.160403			
Final weighted performance:	0.0871788			

approaches to the problem of combined classifier optimisation, producing substantially better classification performance than the more conventional weight optimisation, albeit at the expense of computation time. That method number 5 (the sequential backward selection algorithm applied to the individually pre-optimised classifier feature sets) also produced some performance improvement over the "pre-optimisation" technique alone (albeit to a lesser degree than weight-optimisation) indicates that we have still not, in preferentially opting for method 4, achieved the optimally performing feature set appropriate to the classifier combination. A technique of alternating forward and backward feature selection passes applied to the pre-optimised feature sets should, to a greater or lesser extent, combine the best of the two differing mechanisms of optimisation (to elaborate: the mechanisms of estimation error reduction attributable to redundant pattern space dimensionality in the case of backward selection, and complementarity of feature information in the case of forward selection). This, along with more complex mechanisms of floating feature selection, remains for further investigation.

We note also that the two disparate methods of combined classifier optimisation that the investigation divides into, namely; weight optimisation and feature-set optimisation, are in no way mutually exclusive. Notwithstanding the parallel format of our presentation of the two techniques for the purposes of comparison and contrast, it is perfectly possibly, without significant addition to the execution time, to apply weight optimisation to the classifier/feature set combination obtained by the prior application of method 4. Thus a further performance enhancement would be expected, the optimisation methodology for the two optimisation techniques being of an entirely distinct nature.

There is a further, more fundamental level at which the two differing techniques might be integrated; rather than the two being applied the consecutive manner set out above, we might instead include weight optimisation immediately prior to the determination of the criterion function for inclusion of individual features in method 4 via an additional series of sub-iterations. Within such a procedure the finite operation time of the weight optimisation would become far more apparent, being repeated at every iteration of the feature selection algo-

rithm, but only to the extent of a fractional increase in the total execution time. This then, along with the aforementioned possibility of additional alternating backward and forward feature selection passes, would appear to represent, to the extent that the scope of the current investigation allows, the most promising direction for future techniques of combined-classifier optimisation to progress.

We end with an observation previously alluded to, namely; that there exists within the sequential forward selection techniques employed above, the generalised tendency for the combined criterion function to favour the addition of features to those classifiers with a pre-existing feature-set, rather to those classifiers as yet without any features attributed, up to the point at which further features increase the misclassification rate for the classifier so favoured. A theory as to the origin of this effect is given below:

6 Discussion

In attempting to establish why, particularly, it is that features complement each other to such a greater extent when contained within a single classifier than when distributed over several classifiers, we might envisage the problem metaphorically in terms of one-dimensional projections of a multi-dimensional pattern space (a total of two dimensions chosen for simplicity throughout the following). We know from the theory of Radon transforms that it is possible to reconstruct a two-dimensional pattern space from one-dimensional line integrals taken at various angles and intervals across that space only if the angular sampling of these lines matches the linear sampling; there is not sufficient information contained within the line integrals for reconstruction of that pattern space for the case in which the linear resolution greatly exceeds the angular resolution. This, however, is exactly the situation that occurs when single features of a pattern space are considered in isolation within separate classifiers; the act of obtaining a single feature for inclusion in a specific classifier is, in effect, to integrate linearly across the superfluous dimensions of that feature space. In the scenario we have outlined, when considering only a total of two feature-space dimensions, the angular samples of the pattern space exist at only two points, namely; the perpendicular axes of the pattern space. Now, the linear resolution is as great as the number of samples in the space, which for our investigation is of the order of 1000; clearly, then, this number is far in excess of the angular sample rate of 2. Therefore, even for classifiers that obtain extremely good classification performance on the two features considered independently, there can be no conceivable method of classifier combination that can recover all of the information that dictates the multi-dimensional morphology of the class structures of the pattern-space. The two feature dimensions when contained within a single classifier, however, have only the limitations associated with the sample size and the classifier itself in determining this morphology. One may therefore see that, once a feature has been allocated to a classifier on the basis of classification performance alone, further feature additions to the same classifier, if made on the same basis within a sequential forward selection scenario, will almost invariably follow; all non-exhaustive selection algorithms that treat combined classifiers

and employ nested feature sets in any manner will almost invariably, and to a similar degree, exhibit this effect.

7 Acknowledgement

This research was carried out at the University of Surrey, UK, supported by, and within the framework of, EPSRC research grant number GR/M61320.

References

1. R A Jacobs, Methods for combining experts' probability assessments, Neural Computation, 3, pp 79-87, 1991
2. J. Kittler, M. Hatef, R.P.W. Duin, and J. Matas, On combining classifiers, IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 20, no. 3, 1998, 226-239
3. L. Lam and C.Y. Suen, Optimal combinations of pattern classifiers, Pattern Recognition Letters, vol. 16, no. 9, 1995, 945-954.
4. A F R Rahman and M C Fairhurst, An evaluation of multi-expert configurations for the recognition of handwritten numerals, Pattern Recognition Letters, 31, pp 1255-1273, 1998
5. A F R Rahman and M C Fairhurst, A new hybrid approach in combining multiple experts to recognise handwritten numerals, Pattern Recognition Letters, 18, pp 781-790, 1997
6. K Woods, W P Kegelmeyer and K Bowyer, Combination of multiple classifiers using local accuracy estimates, IEEE Trans. Pattern Analysis and Machine Intelligence, 19, pp 405-410, 1997
7. A. Hojjatoleslami and J. Kittler, Strategies for weighted combination of classifiers employing shared and distinct representations, International Conference on Pattern Recognition, pp 338-340, 1998
8. P. A. Devijver and J. Kittler, Pattern Recognition: A Statistical Approach, 1982, Prentice-Hall International, Inc. ISBN 0-13-654236-0.