Multiple Expert System Design by Combined Feature Selection and Probability Level Fusion

F. M. Alkoot  
C. V. S. S. P.  
School of E.E.I.T.M.  
University of Surrey  
Guildford GU2 5XH  
United Kingdom  
F.Alkoot@eim.surrey.ac.uk  

J. Kittler  
C. V. S. S. P.  
School of E.E.I.T.M.  
University of Surrey  
Guildford GU2 5XH  
United Kingdom  
J.Kittler@eim.surrey.ac.uk

Abstract - We propose a novel design philosophy for expert fusion by taking the view that the design of individual experts and fusion cannot be solved in isolation. Each expert is constructed as part of the global design of a final multiple expert system. The design process involves jointly adding new experts to the multiple expert architecture and adding new features to each of the experts in the architecture.

We evaluate the performance of different fusion strategies ranging from linear untrainable strategies like Sum and Modified Product to linear and nonlinear trainable strategies like logistic regression, single layer perceptron and radial basis function classifier.

We investigate two distinct design strategies which we refer to as parallel and serial. In both cases we show that the proposed integrated design approach leads to improved performance.

Keywords: Expert Fusion, Feature Selection, Gaussian Classifier, Nearest Neighbour Classifier, Neural Networks.

1 Introduction

The worthwhile improvements in performance that can be gained from the fusion of multiple classifiers have been widely documented [1, 2, 3, 4]. The surprising benefits that can be accrued from the combination of multiple classifier designs over the best single expert has motivated a substantial research community to develop and investigate diverse approaches to multiple expert fusion [5, 6, 7, 8]. These range from simple combination rules which do not require any training [9, 10], to sophisticated fusion mechanisms which can cope with unequal expert strengths, can even dynamically adapt to data [11, 12, 13, 14, 15, 16].

In spite of the difference in the underlying fusion principles, almost all the techniques proposed in the literature share the same premise, namely that the experts involved in fusion have already been designed using conventional methodology. This implies that the fusion is considered as an after thought, addressing the question how the experts already existing can be combined in the most effective way. The design of individual experts which involves the choice of classifier (e.g. Gaussian, k-nearest neighbour, radial basis function, multi-layer perceptron) and the corresponding feature space is carried out using the standard criterion of classification performance the experts would achieve individually. It is at the stage of fusion, when a classifier combination scheme is being devised, that the merits of the individual experts are assessed and the expert opinions combined accordingly. In some cases such a design approach may adhere to a deliberate policy of breaking up a complex design problem into simpler tasks. However, in most cases it will simply be a reflection of the incremental development of the design methodology.

It is generally believed that the fusion of multiple experts is beneficial if they voice independent opinions. Methods to build uncorrelated experts can mainly be grouped in two. One is based upon diversifying the training set, as in bagging and boosting [17, 18], the other would be to diversify the architecture of the component experts [19]. This is achieved by altering the initial weights of several neural network experts, or to use a different architecture for each expert such as a neural network combined with gaussian or K-nearest neighbour classifiers.

While it may be largely true that combining independent experts leads to performance improvement, it is very easy to demonstrate that even statistically dependent experts can be combined profitably. This is why the problem of choosing a suitable fusion strategy is so difficult. As we do not know the characteristics of the respective experts, it is far from clear which strategy should be used as the fusion schemes which are able to capitalise on expert independence will not necessarily fare well when the experts outputs are dependent.

In our work we adopt a completely novel design philosophy. We take the view that the design of individual experts and fusion cannot be solved in isolation. This premise leads to a completely different design methodology whereby each expert is constructed as part of the global design of a final multiple expert system. As the design of component classifiers is optimised using a common performance criterion, that is the error probability of the multiple expert system, it is reasonable to expect that the final design will be at least as good as the fusion of individually designed experts and hopefully much better.

The design process involves jointly adding new experts to the multiple expert architecture and adding new features to each of the experts in the architecture. The feature selection problem itself is of combinatorial complexity and
it is clear that the optimisation over different architectures and individual experts in each architecture will be computationally explosive. For this reason in this initial study we use only the simplistic sequential forward feature selection method to build the individual experts and the fusion system.

We investigate two distinct design strategies which we refer to as parallel and serial. In the parallel approach we utilise a pool of features by building a number of experts simultaneously by distributing the features available to the experts in a "card dealing" manner. In this approach the dimensionality of the feature spaces in which the experts operate increases in a balanced way and also the expert strengths are reasonably balanced. In contrast, in the serial approach any expert is allowed to absorb new features as long as the system performance continues to improve. This inevitably means that the first experts take better and more features than the experts added later, leading to imbalance both in feature space dimensionality and individual expert performance. We also explore the merits of using different types of classifiers and their mixing in the multiple expert system design.

The different designs are also distinguished by the choice of fusion strategy. We investigate the ubiquitous sum rule [9, 10] and the recently proposed modified product [20] which do not require any training. Their performance is compared with trainable linear combination strategies, namely logistic regression [21, 22, 2] and single layer perceptron and with that afforded by a nonlinear combiner realised using a radial basis function classifier.

We illustrate the effectiveness of our method on two classification problems: breast cancer detection [23], and Seismic data interpretation. In both cases we show that the proposed integrated design approach leads to improved performance. The fusion strategy yielding the best results appears to be different for the two applications. It would appear that in the breast cancer case the measurements are quite independent as reflected in the sustained performance of the single experts in the full observation space, whereas the Seismic data contains redundant features. For the latter application the different experts are likely to be highly correlated and logistic regression secures optimal fusion. In the former case the expert outputs are more likely to be independent and this explains why the sum combination rule outperforms other contenders. As expert correlation can easily be measured these cases could be identified automatically and the fusion strategy selected accordingly.

The paper is organised as follows. In the next section we will introduce the proposed method of building multiple expert systems. In section 3 the details of combining the classification experts used and of the inherent fusion strategies are presented. The results of experiments on two data sets are discussed in sections 4 and 5 respectively. The paper is drawn to conclusion in section 6.

2 The Proposed Method

In our multiple expert system, in order to achieve diversity, we propose a distinct set of features for each expert. Hence experts are designed by assigning features to them. At the end of the design process each expert will hold a distinct set of features. We propose two methods of building the experts of the system, parallel and serial.

2.1 The Parallel System

In the parallel method, at any stage of the design an expert is allowed to take the feature that will deliver the best system performance. In order to encourage a balance in the performance of individual experts each expert is allowed to take only one feature at a time. The number of experts is a system parameter that is specified at the beginning of the design process. When a feature is introduced to an expert the system classification rate is estimated using a validation set. The feature that optimises the system performance is assigned to the expert. When the first feature is introduced the system consists of only one expert, hence no fusion occurs. The second feature is introduced to a new expert and the feature that results in the maximum system performance is selected for this second expert. When all experts are assigned a feature we go back to the first expert and assign it a second feature. This process continues until all features are selected or the termination condition is reached. A possible termination condition is when the addition of any of the remaining features fails to improve the system performance. The parallel method allows best features to be distributed among many experts. All experts use similar criterion functions. Therefore we expect the experts to have approximately equal strength.

2.2 The Serial System

In the serial method the first expert is allowed to take all the features that it needs to achieve maximum performance. Next, if any feature remains, a second expert is built from the remaining features. It is now, when a feature is added to the second expert, that we observe the effect of fusion. The process of building experts continues until all features are selected or a termination condition is reached.

This method allows the first expert to take the best features. Therefore the subsequent experts will be weaker. This will lead into a more diverse set of experts, as far as their relative strength is concerned, compared to the experts in the parallel system. Logistic regression or any kind of weighted fusion is expected to perform better than the simple Sum strategy.

In the serial method the number of experts is variable and depends on the number of features taken by the experts. It could vary from a single expert up to a maximum equal to the number of features.

3 Experimental Methodology

3.1 Multiple expert systems

We experiment with four types of classification experts generating the class posteriori probabilities in different ways, namely the gaussian, K-nearest neighbour (KNN), nearest neighbour (NN) and radial basis function experts. When using KNN experts we set \( k = \sqrt{N} \), where \( N \) is the number of training samples.
The multiple expert systems are built in two different ways so as to create either homogeneous systems or heterogeneous systems. Any homogeneous system is built from one type of expert only.

In contrast a heterogeneous multiple expert system contains a mixture of all of the above experts. The heterogeneous systems are constructed using as the first four experts the gaussian, k-nearest neighbour, l-nearest neighbour and radial basis function classifiers respectively. All additional experts are gaussian classifiers. This heterogeneous system is referred to in the following tables as Mult4. Another heterogeneous system which excludes the radial basis function classifier is referred to as Mult3.

The aim of a heterogeneous design is to end up with uncorrelated and diverse experts as combining uncorrelated experts results in higher correct classification rates. In homogeneous systems we achieve diversity only through the use of different features. By virtue of heterogeneity we introduce another factor that increases diversity.

We have used a simple forward feature selection method [24] to select the best feature(s) for each expert. More sophisticated feature selection methods [25], if used, may lead to better results.

3.2 Fusion Strategies

One of the design specifications of the proposed methodology for multiple expert system construction is the fusion rule. In our approach we focus on decision probability level fusion. We assume that our experts deliver a soft opinion for each possible class expressed in terms of the aposteriori probability for each class as required to ensure that the normalised quantities always sum up to one.

The linear fusion rules established by training on the validation set are of the form

\[ S_i = \sum_{j=0}^{R} \beta_{ji} s_{ji} \]

where \( \beta_{ji}, j = 1, R, \) the components of the parameter vector \( \beta_j \), are the mixing parameters of the scores \( s_{ji} \) and \( \beta_{0i} \) represents the offset-set with \( \beta_{0i} = 1 \). The following linear rules have been studied:

**Nearest Mean Fusion**[26] The linear discriminant function corresponding to this decision rule is given by

\[ \beta_j = \frac{X^T Y_i}{||X^T Y_i||} \]

where \( X \) is the data matrix of scores for the elements in the training set

\[ X = \begin{bmatrix} s_{1i}^1 & \cdots & s_{1i}^N \\ s_{2i}^1 & \cdots & s_{2i}^N \\ \vdots & \ddots & \vdots \\ s_{ni}^1 & \cdots & s_{ni}^N \end{bmatrix} \]

and the vector of target responses \( Y_i \) has elements \( \frac{1}{N} \) for patterns from class \( \omega_i \) and \( \frac{1}{N-1} \) for all the other patterns. \( \beta_{0i} \) is set so that a point on the boundary maps to \( \frac{1}{2} \). This fusion strategy will be referred to as NrMean in the tables.

**Linear Regression** By augmenting the vector of scores by an \((R + 1)st\) component \( s_{0i} \) set to one and the parameter vector by the corresponding element \( \beta_{0i} \) we can find the least squares fit to the binary target output. The solution is given by

\[ \beta_j = (X^T X)^{-1} X^T Y_i \]

where

\[ X = \begin{bmatrix} 1 & s_{1i}^1 & \cdots & s_{1i}^N \\ 1 & s_{2i}^1 & \cdots & s_{2i}^N \\ \vdots & \vdots & \ddots & \vdots \\ 1 & s_{ni}^1 & \cdots & s_{ni}^N \end{bmatrix} \]

and the vector of target responses \( Y_i \) has its elements set to one for patterns from class \( \omega_i \) else zero. This regression will be referred to as LnReg in the tables.

**Logistic Regression** Under the assumption that the expert outputs are conditionally distributed according to a multivariate normal, they can be combined using logistic regression with weighting parameters

\[ \beta_d = (\mu_i - \mu_0)^T \Sigma^{-1} \]

and an off-set

\[ \beta_{0i} = \log\left( \frac{N_i}{N-N_i} \right) - 0.5(\mu_i - \mu_0)^T \Sigma^{-1}(\mu_i + \mu_0) \]
where $\mu_i$ is the mean vector of expert scores for patterns in class $\omega_i$ and $\mu_0$ is the vector of mean scores for all the other classes. $\Sigma$ is the weighted average covariance matrix of the respective score populations. This strategy will be referred to as LogReg in the tables.

Table 1: Performance comparison of fusion strategies for the parallel system design. Using BCW data for the termination condition.

<table>
<thead>
<tr>
<th>Fusion strategy</th>
<th>Gauss</th>
<th>KNN</th>
<th>1NN</th>
<th>Mult4</th>
<th>Mult3</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sum</td>
<td>93.32</td>
<td>93.16</td>
<td>90.95</td>
<td>93.12</td>
<td>92.92</td>
</tr>
<tr>
<td>Mprod</td>
<td>93.16</td>
<td>93.36</td>
<td>90.95</td>
<td>93.12</td>
<td>92.92</td>
</tr>
<tr>
<td>NrMean</td>
<td>91.52</td>
<td>69.38</td>
<td>70.72</td>
<td>80.27</td>
<td>79.50</td>
</tr>
<tr>
<td>LnrReg</td>
<td>92.92</td>
<td>93.56</td>
<td>89.38</td>
<td>94.22</td>
<td>94.59</td>
</tr>
<tr>
<td>LogReg</td>
<td>92.92</td>
<td>93.56</td>
<td>88.48</td>
<td>94.26</td>
<td>94.59</td>
</tr>
<tr>
<td>Percp</td>
<td>91.72</td>
<td>91.62</td>
<td>91.62</td>
<td>91.45</td>
<td>91.45</td>
</tr>
<tr>
<td>PercpSm</td>
<td>92.49</td>
<td>92.35</td>
<td>91.49</td>
<td>92.79</td>
<td>92.05</td>
</tr>
<tr>
<td>Radbas</td>
<td>64.78</td>
<td>84.07</td>
<td>91.42</td>
<td>71.45</td>
<td>71.45</td>
</tr>
<tr>
<td>RadbasSm</td>
<td>92.02</td>
<td>86.91</td>
<td>90.58</td>
<td>90.69</td>
<td>91.59</td>
</tr>
<tr>
<td>Singl xprt</td>
<td>92.55</td>
<td>93.86</td>
<td>91.05</td>
<td>-</td>
<td>-</td>
</tr>
</tbody>
</table>

Table 2: Performance comparison of fusion strategies for the parallel system design, using BCW data. Considering all features.

<table>
<thead>
<tr>
<th>Fusion strategy</th>
<th>Gauss</th>
<th>KNN</th>
<th>1NN</th>
<th>Mult4</th>
<th>Mult3</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sum</td>
<td>96.63</td>
<td>95.83</td>
<td>92.72</td>
<td>95.66</td>
<td>96.26</td>
</tr>
<tr>
<td>Mprod</td>
<td>95.66</td>
<td>95.83</td>
<td>92.72</td>
<td>95.56</td>
<td>96.06</td>
</tr>
<tr>
<td>NrMean</td>
<td>93.99</td>
<td>85.91</td>
<td>83.57</td>
<td>88.25</td>
<td>88.08</td>
</tr>
<tr>
<td>LnrReg</td>
<td>95.23</td>
<td>95.76</td>
<td>95.09</td>
<td>95.43</td>
<td>95.13</td>
</tr>
<tr>
<td>LogReg</td>
<td>95.23</td>
<td>95.76</td>
<td>89.08</td>
<td>95.36</td>
<td>95.13</td>
</tr>
<tr>
<td>Percp</td>
<td>91.89</td>
<td>95.13</td>
<td>93.36</td>
<td>93.29</td>
<td>94.99</td>
</tr>
<tr>
<td>PercpSm</td>
<td>93.99</td>
<td>95.19</td>
<td>92.32</td>
<td>91.89</td>
<td>93.96</td>
</tr>
<tr>
<td>Radbas</td>
<td>90.18</td>
<td>94.83</td>
<td>94.16</td>
<td>86.64</td>
<td>87.45</td>
</tr>
<tr>
<td>RadbasSm</td>
<td>89.68</td>
<td>93.56</td>
<td>92.92</td>
<td>88.92</td>
<td>77.80</td>
</tr>
<tr>
<td>Singl xprt</td>
<td>92.29</td>
<td>95.33</td>
<td>95.06</td>
<td>-</td>
<td>-</td>
</tr>
</tbody>
</table>

3.3 Data sets

Breast Cancer Wisconsin [23] (BCW): This two class data set consists of 699 nine feature samples. The training set consists of 50 randomly selected samples. The number of samples per class are taken such that their proportion is equal to their original proportion. Using the rest of the samples we construct the validation set in a similar manner. The remaining samples constitute the test set.

Seismic data: This data set was obtained from Shell for research purposes at CVSSP, University of Surrey. It contains three classes. The data is divided into two sets L0 and T0. L0 contains 25229 samples while T0 contains 17560 samples. Samples of both files have 25 features. We randomly sample 500 samples from each class to build each of the training, validation and test sets. Samples from set L0 are used for the training and validation sets, while the test set is constructed from T0.

4 Results

Classification rates are indicated in tables 1-8. Whenever the performance is higher than the highest single expert it is highlighted in bold. Results are also plotted in order to obtain a general overview.

In all figures the symbols ♦, O and □ represent the performance of gaussian, KNN and NN experts, respectively. The △ and ★ represent the heterogeneous system using the four types of experts (Mult4) mentioned in section 3 and the heterogeneous system using three types of experts excluding the radial basis function (Mult3), respectively. Missing values imply very low performance. The correct classification rates are compared to the single expert, (kNN, Gaussian and NN) shown by horizontal lines.

4.1 The Parallel System

For BCW data, the design obtained using the termination condition method achieves the best performance using a single kNN expert. Our proposed method achieves a better performance only for the heterogeneous system if fused using regression methods as shown in Table 1. Note that the
parallel method improves the gaussian expert.

More impressive performance gains are made when we use all the features and stop the system construction when the last feature is introduced. An exception is when NN experts are fused. The single NN expert performs better than the fused NN experts.

The largest improvement is achieved when the gaussian function is used. Actually using Sum or MProduct to fuse gaussian experts we get the best overall classification rate, as shown in Table 2.

A heterogeneous expert system gives good results but not optimum. Mostly its performance falls between the performance of Gaussian, KNN and NN based systems. If the heterogeneous system is fused using regression methods we achieve optimum performance.

For the Seismic data nearest mean fusion of gaussian experts yields the best results. Most fusion methods fail to outperform the single KNN expert. Fusing NN experts using the perceptron achieves an optimum performance. The perceptron performs well when all features are used.

In general applying this method to gaussian experts is successful but less so when NN experts are involved. When fusing using Sum, MProduct or Logistic regression, Gaussian based systems exhibit the best performance.

### 4.2 The Serial System

For the BCW data, considering the termination condition case, the serial method of multiple expert system design improves on the single expert for all types of experts, especially when Sum, MProduct or regression strategies are used to fuse the experts. Overall the highest performance achieved is that of the KNN single expert as seen from Table 5, with an insignificant improvement when regression strategies are used on kNN or the heterogeneous system.

When all features are used we again notice from Table 6 that the gaussian experts obtain the best performance if fused using Sum or MProduct. Fusing kNN or NN does not lead to an improvement over the single expert.

In general, when using all features, parallel and serial designs respond with the highest performance if Sum or MProduct are used to combine the experts.

Using the termination condition method the heterogeneous system yields good results although none of the fusion methods yield optimum performance, except for regression strategies. When all features are used in a heterogeneous system Sum and MProduct yield a very good system performance that is better than any type of single expert.

For the Seismic data we notice that the serial method improves on the single expert if fused using regression. The perceptron yields a very high performance consistently using any type of expert, although it is not optimal.
Table 5: Performance comparison of fusion strategies for the serial system design. Using BCW data for the termination condition.

<table>
<thead>
<tr>
<th>Fusion strategy</th>
<th>Gauss</th>
<th>KNN</th>
<th>1NN</th>
<th>Mult4</th>
<th>Mult3</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sum</td>
<td>93.76</td>
<td>93.86</td>
<td>91.79</td>
<td>93.26</td>
<td>93.26</td>
</tr>
<tr>
<td>Mprod</td>
<td>93.72</td>
<td>93.86</td>
<td>91.79</td>
<td>93.26</td>
<td>93.26</td>
</tr>
<tr>
<td>NrMean</td>
<td>91.09</td>
<td>93.86</td>
<td>91.05</td>
<td>91.82</td>
<td>91.85</td>
</tr>
<tr>
<td>LnrReg</td>
<td>93.62</td>
<td>93.89</td>
<td>91.12</td>
<td>93.66</td>
<td>93.89</td>
</tr>
<tr>
<td>LogReg</td>
<td>93.59</td>
<td>93.89</td>
<td>91.12</td>
<td>93.66</td>
<td>93.89</td>
</tr>
<tr>
<td>Percp</td>
<td>90.28</td>
<td>93.86</td>
<td>90.92</td>
<td>91.12</td>
<td>90.79</td>
</tr>
<tr>
<td>PercpSm</td>
<td>92.65</td>
<td>93.39</td>
<td>91.79</td>
<td>92.29</td>
<td>92.29</td>
</tr>
<tr>
<td>Radbas</td>
<td>73.66</td>
<td>90.79</td>
<td>91.59</td>
<td>80.10</td>
<td>80.10</td>
</tr>
<tr>
<td>RadbasSm</td>
<td>90.38</td>
<td>93.69</td>
<td>92.17</td>
<td>93.22</td>
<td>93.22</td>
</tr>
<tr>
<td>Singl xprt</td>
<td>92.55</td>
<td>93.86</td>
<td>91.05</td>
<td>-</td>
<td>-</td>
</tr>
</tbody>
</table>

Table 6: Performance comparison of fusion strategies for the serial system design using BCW data. Considering all features.

<table>
<thead>
<tr>
<th>Fusion strategy</th>
<th>Gauss</th>
<th>KNN</th>
<th>1NN</th>
<th>Mult4</th>
<th>Mult3</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sum</td>
<td>96.26</td>
<td>94.06</td>
<td>91.89</td>
<td>95.93</td>
<td>95.63</td>
</tr>
<tr>
<td>Mprod</td>
<td>96.39</td>
<td>94.19</td>
<td>91.89</td>
<td>95.86</td>
<td>95.63</td>
</tr>
<tr>
<td>Log</td>
<td>89.72</td>
<td>80.30</td>
<td>83.21</td>
<td>87.68</td>
<td>87.61</td>
</tr>
<tr>
<td>LnrReg</td>
<td>94.26</td>
<td>93.76</td>
<td>91.05</td>
<td>94.39</td>
<td>94.83</td>
</tr>
<tr>
<td>LogReg</td>
<td>94.26</td>
<td>93.76</td>
<td>69.85</td>
<td>94.39</td>
<td>94.83</td>
</tr>
<tr>
<td>Percp</td>
<td>90.05</td>
<td>94.42</td>
<td>91.19</td>
<td>92.69</td>
<td>92.92</td>
</tr>
<tr>
<td>PercpSm</td>
<td>89.85</td>
<td>94.62</td>
<td>90.25</td>
<td>90.62</td>
<td>91.29</td>
</tr>
<tr>
<td>Radbas</td>
<td>91.72</td>
<td>91.55</td>
<td>93.09</td>
<td>92.99</td>
<td>93.06</td>
</tr>
<tr>
<td>RadbasSm</td>
<td>91.49</td>
<td>92.76</td>
<td>92.43</td>
<td>92.86</td>
<td>91.72</td>
</tr>
<tr>
<td>Singl xprt</td>
<td>92.29</td>
<td>95.33</td>
<td>95.06</td>
<td>-</td>
<td>-</td>
</tr>
</tbody>
</table>

5 Discussion

It seems that when using the Seismic data the systems over train and hence can not generalise well during testing. When the test and training sets are both from one file, for example L0, we get a very high classification rate of 99.97 using the Sum rule in a parallel gaussian system.

The best performance on BCW that we achieve is when all features are considered using gaussian experts and the experts are fused with Sum and MProduct. Both serial and parallel systems obtain rates higher than 96.2 compared to a single experts performance of 92.29 as shown in Tables 2 and 6. Skalak [27] on BCW achieves 96.2 using NN, 96.4 using his single prototype method and 96.8 using his two prototype method. As our NN rate is lower than his due to working with a different training set we could expect with our method to achieve better than his best performance.

5.1 Comparing Architectures

In general Serial is better on KNN and NN while parallel is better on gaussian although both design methods improve gaussian experts. Parallel sometimes improves kNN.

Serial does not degrade much if all features are used. In the serial case assigning new features to new experts does not degrade old experts and keeps the system resilient to the peaking phenomenon. The parallel design degrades when new features are added.

When fusing kNN or NN by serial design, regression methods do better than the single expert. In general we do not benefit from fusing KNN or NN experts. Few exceptions exist like when the serial system is used on the Seismic data. The reason could be due to the ability of the serial system to build un-correlated experts even at the presence of redundant features. It is more successful than the parallel system at building uncorrelated experts.

Gaussian experts are weaker than KNN, hence even in the presence of redundant features each expert is different from the rest and hence fusion is successful on both data types and both systems.

5.2 Comparing fusion strategies

Among all the rules we notice that Sum, MProduct and Linear Regression were most successful. Linear Regression was best on the Seismic data when the serial system was used. Sum and MProduct were best on the BCW data when either system was used. But their superiority was more ob-
Table 7: Performance comparison of fusion strategies for the serial system design. Using Seismic data for the termination condition.

<table>
<thead>
<tr>
<th>Fusion Strategy</th>
<th>Gauss</th>
<th>KNN</th>
<th>1NN</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sum</td>
<td>77.61</td>
<td>80.65</td>
<td><strong>82.01</strong></td>
</tr>
<tr>
<td>Mprod</td>
<td>78.08</td>
<td><strong>81.96</strong></td>
<td><strong>82.01</strong></td>
</tr>
<tr>
<td>NrMean</td>
<td>73.25</td>
<td><strong>82.19</strong></td>
<td><strong>82.47</strong></td>
</tr>
<tr>
<td>LnReg</td>
<td><strong>84.00</strong></td>
<td>84.07</td>
<td><strong>81.63</strong></td>
</tr>
<tr>
<td>LogReg</td>
<td>78.12</td>
<td><strong>83.20</strong></td>
<td><strong>82.32</strong></td>
</tr>
<tr>
<td>Percp</td>
<td>78.21</td>
<td><strong>84.01</strong></td>
<td>79.73</td>
</tr>
<tr>
<td>Singl xpert</td>
<td>64.65</td>
<td>81.23</td>
<td>80.36</td>
</tr>
</tbody>
</table>

Table 8: Performance comparison of fusion strategies for the serial system design using Seismic data. Considering all features.

<table>
<thead>
<tr>
<th>Fusion Strategy</th>
<th>Gauss</th>
<th>KNN</th>
<th>1NN</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sum</td>
<td>77.60</td>
<td>76.95</td>
<td>79.21</td>
</tr>
<tr>
<td>Mprod</td>
<td>78.72</td>
<td>78.20</td>
<td>77.76</td>
</tr>
<tr>
<td>NrMean</td>
<td>79.99</td>
<td><strong>84.97</strong></td>
<td>79.88</td>
</tr>
<tr>
<td>LnReg</td>
<td><strong>82.89</strong></td>
<td>85.57</td>
<td>52.27</td>
</tr>
<tr>
<td>LogReg</td>
<td><strong>82.20</strong></td>
<td>82.95</td>
<td>33.33</td>
</tr>
<tr>
<td>Percp</td>
<td><strong>81.01</strong></td>
<td>83.77</td>
<td><strong>83.08</strong></td>
</tr>
<tr>
<td>Singl xpert</td>
<td>25.71</td>
<td>80.31</td>
<td>79.19</td>
</tr>
</tbody>
</table>

vious in the parallel system. Logistic regression was close
to but never better than Linear Regression.

When compared to Sum, MProduct was better when the
serial system was used, while Sum yielded better performance when the parallel system was used.

When using gaussian experts on Seismic data we noticed
that Sum and MProduct performed better on parallel than
on serial for the designs produced by the termination con-
dition. In the parallel case the performance is close to the
regression methods. The reason NN does not show a good
performance using regression methods is that the covari-
ance matrix in equations 6 and 8 becomes singular.

5.3 The heterogeneous system

We notice that the heterogeneous system, if fused using re-
gression methods, yields very good results. For BCW data
it achieves an optimum performance in a parallel system.
The combination of a heterogeneous system and regression
fusion yield results better than those produced by any het-
erogeneous system using any other fusion method or a ho-
momogeneous system using regression fusion methods. An
exception is when all features are used in a gaussian homo-
geneous system if fused using Sum or MProduct.

6 Conclusion

We proposed a novel design philosophy for expert fusion
by taking the view that the design of individual experts and
fusion cannot be solved in isolation. Each expert was con-
structed as part of the global design of a final multiple ex-
pert system. We investigated two distinct design strategies
which we referred to as parallel and serial. In both cases we
showed that the proposed integrated design approach led to
improved performance.

We evaluated the performance of different fusion strate-
gies ranging from linear untrainable strategies like Sum and
Modified Product to linear and nonlinear trainable strate-
gies like logistic regression, single layer perceptron and ra-
dial basis function classifier.

The good performance of Sum and MProduct in both
systems when BCW was used indicates that BCW features
are distinct and the experts produced have equal powers.
More sophisticated fusion methods overfit the validation
data and hence they perform worse than Sum and MProduct
on the test data.

It seems that serial designs are best when features are
distinct while parallel are good in the presence of the
peaking phenomenon as in the case of the Seismic data. In
the case of weak experts both serial and parallel improve
on the single expert. When strong experts exist serial can
uncorrelate them but a more sophisticated fusion method is
needed, like Logistic regression. Sum and MProduct can
be useful if the resulting uncorrelated experts have similar
Acknowledgement: F. Alkoot would like to thank the Islamic Development Bank and the Public Authority for Applied Education and Training of Kuwait for financially supporting his education towards the Ph.D. The work was also partially supported by EPSRC research grant GR/M61320.

References


