Introduction

• Ensembles of classifiers has been widely used to:
  – Reduce model uncertainty.
  – Improve generalization performance.

• Good ensemble consists of:
  – Good classifiers.
  – Make errors on different parts of the feature space.
Tree Fundamental Reasons why an Ensemble may Work Better Than a Single Classifier

[Dietterich 2000]

Statistical

• The Statistical Problem Arises when the amount of training data available is too small compared to the size of the hypothesis space.

• … The learning algorithm can find different hypothesis in $H$ that all give the same accuracy on the training data.
Computational

• Many learning algorithms work by performing some form of *local search* that may get stuck in *local optima*
  – Ex: neural network & gradient descent

Representational

• In most application of machine learning, the *true function* $f$ cannot be represented by any of the hypotheses in $H$. 
Methods For Ensembles

• Classical methods:
  – **Bagging** [Breiman’96]
  – **Boosting** [Freund’97]
  – **Random subspace** [Ho’98]
    • Varies the subsets of features.
  – The literature has shown that by varying the subsets of features used by each member of the ensemble should help to promote diversity.

Methods For Ensembles

• GA-based methods:
  – Single GA.
  – Varies the subsets of features by performing feature selection.
  – Usually produce only one ensemble [Optiz99].
  – Must combine multiple objective functions into one global function.
The Proposed Method

- Based on a hierarchical multi-objective GA.
  - 1\textsuperscript{st} level performs features selection.
    - Finds a set of good (diverse) classifiers.
  - 2\textsuperscript{nd} combines those classifiers.
    - Maximizing the generalization power of the ensemble and a measure of diversity.
    - Produces a set of ensembles.
1\textsuperscript{st} Level – Feature Selection

- Genetic algorithms:
  - Effective in rapid global search of large and poorly understood spaces.
  - Attractive approach to deal with multi-criterion optimisation.
- Wrapper and filter.
- Why wrapper instead of filter:
  - It takes into account the learning algorithm, so that representation biases of the classifier are considered.
- Modified wrapper.
  - Sensitivity analysis and Neural Nets [Emmanouilidis00].
  - Validation set to avoid overfitting.

Multi-objective Optimization Problem

- It consists of a number of objectives which are associated with a number of inequality and equality constraints.
- Solutions can be expressed in terms of non-dominated points.
  - A solution is dominant over another only if it has superior performance in all criteria.
  - All non-dominated solutions compose the Pareto-optimal front.
Multi-objective Optimization Problem

- Classical approach (weighted sum).
  - Multiple objectives are combined into a single and parameterized objective.
    \[ F(x) = \sum_{i=1}^{N} \omega_i f_i(x) \]
  - Drawbacks:
    - Scaling.
    - Dependence of the weights.
    - One solution.

Multi-objective GA
Multi-objective GA

- Pareto-based approach [Goldberg89]:
  - It uses Pareto dominance in order to determine the reproduction probability of each individual.
  - Fitness sharing:
    - Individuals in a particular niche have to share their fitness in order to maintain the diversity.
    - The more individuals are located in the neighbourhood of a certain individual, the more its fitness value is degraded.

Non-dominated Sorting GA (NSGA)

- [Srinivas & Deb 95].
- Ranking by fronts.
- It converges close to the Pareto-optimal front.
- Sharing is achieved by:

\[
Sh(d(i, j)) = \begin{cases} 
1 - \frac{d(i, j)}{\sigma_{share}} & \text{if } d(i, j) < \sigma_{share} \\
0 & \text{Otherwise}
\end{cases}
\]

where \( \sigma_{share} \approx \frac{0.5}{\sqrt{q}} \)  \( p = \text{number of decision variables} \)  \( q \approx 10 \)
Flow Chart Of The Methodology

Methodology

- NSGA.
  - Bit representation, one-point crossover, bit-flip mutation, and elitism.
- Fitness evaluation:
  \[ f_1 = \text{Number of selected features}. \]
  \[ f_2 = \text{Error rate of the classifier}. \]
Methodology

• Sensitivity analysis [Utans & Moody’91].
  – It substitutes the unselected features by their averages, which are computed on the training set.
  – It avoids training the neural network for each different subset of features generated during the search.

Methodology

• Validating the Pareto-optimal front.
  – It points out the solution with better generalization power.
  – Validation set (2):
    • 30,000 samples (hsf_7).
Handwritten Digit Classifiers

- MLP trained with backpropagation.
- Database: NIST SD19:
  - Training set: 195,000 (hsf_0123)
  - Validation set: 28,000 (hsf_0123)
  - Test set: 30,089 (hsf_7)

<table>
<thead>
<tr>
<th>Feature Set</th>
<th>No. of Features</th>
<th>RR %</th>
</tr>
</thead>
<tbody>
<tr>
<td>Distances</td>
<td>[Oh &amp; Suen, 98]</td>
<td>96</td>
</tr>
<tr>
<td>Edge Maps</td>
<td>[Chim et al, 98]</td>
<td>125</td>
</tr>
</tbody>
</table>

Experiments

- Classical approach (monobjective optimization).
  - It presents a premature convergence to a specific region instead of maintaining a diverse population.
Experiments

- Pareto-based approach (multi-objective optimization).
  - It converges close to the Pareto-optimal front.
  - Importance of validating the Pareto-optimal front.

Results Of Feature Selection

- Single-population master-slave GA.
- Beowulf cluster with 17 machines (1.1ghz, 512 RAM).
  - MPI-LAM [http://www.lam-mpi.org/].
  - About 4 hours per experiment.

<table>
<thead>
<tr>
<th>Feature Set</th>
<th>Original Classifier</th>
<th>Optimized Classifier</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Features</td>
<td>RR %</td>
</tr>
<tr>
<td>Concavities</td>
<td>132</td>
<td>99.13</td>
</tr>
<tr>
<td>Distances</td>
<td>96</td>
<td>98.17</td>
</tr>
<tr>
<td>Edge Maps</td>
<td>125</td>
<td>97.04</td>
</tr>
</tbody>
</table>
2nd Level: Finding Ensembles

- To combine the classifiers produced in the previous level to provide a set of powerful ensembles.

- Each gene of the chromosome stands for a classifier of the Pareto generated in the 1st level.
  - If a chromosome has all bits selected, all classifiers of the Pareto will be included in the ensemble.

Types Of Classifiers In The Pareto
Summary Of The 1st Level Classifiers

<table>
<thead>
<tr>
<th>Feature Set</th>
<th>No. of Classifiers</th>
<th>Range of Features</th>
<th>Range of RR (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Concavities</td>
<td>81</td>
<td>24 – 125</td>
<td>90.5 – 99.1</td>
</tr>
<tr>
<td>Distances</td>
<td>54</td>
<td>30 – 84</td>
<td>90.6 – 98.1</td>
</tr>
<tr>
<td>Edge Maps</td>
<td>78</td>
<td>35 – 113</td>
<td>90.5 – 97.0</td>
</tr>
</tbody>
</table>

2nd-level Population

1st Level Pareto

2nd-Level Population

1

2

...n
Objective Functions

- To find the most diverse set of classifiers that brings a good generalization.
  - Maximization of the recognition rate of the ensemble.
  - Maximization of a measure of diversity [Kuncheva’02]:
    - Overlap.
    - Entropy.
    - Ambiguity.

Ambiguity

\[ a_i(x_k) = \frac{1}{k} \left[ V_i(x_k) - \overline{V}(x_k) \right]^2 \]

where \( a_i \) is the ambiguity of the \( i^{th} \) classifier on the example \( x_k \)
while \( V_i \) and \( \overline{V} \) are the \( i^{th} \) classifier and the ensemble predictions respectively.

- It is simply the variance of the ensemble around the mean, and it measures the disagreement among the networks on input \( x \).
Combination Of Classifiers

- It is necessary in order to compute the generalization power of the ensemble.
  - Average.
  - The literature has been shown that it is a simple and effective scheme of combining predictions of the neural networks [Kittler et al, 98].

Ensembles Produced By The 2nd Level

- Concavities
- Distances
- Edge Maps
Concavities

Number of classifiers

Distances

Number of classifiers
Edge Maps

Number of classifiers

Performance of the Ensembles

<table>
<thead>
<tr>
<th>Feature Sets</th>
<th>Number of Classifiers</th>
<th>RR (%) zero-rejection level</th>
<th>Single Classifier</th>
</tr>
</thead>
<tbody>
<tr>
<td>Concavities</td>
<td>4</td>
<td>99.23</td>
<td>99.13</td>
</tr>
<tr>
<td>Distances</td>
<td>4</td>
<td>98.16</td>
<td>98.17</td>
</tr>
<tr>
<td>Edge Maps</td>
<td>7</td>
<td>97.16</td>
<td>97.04</td>
</tr>
</tbody>
</table>

- Same performance when working at zero-rejection level.
- Compelling improvements when working at low error rates
  - Real systems.
Concavities

Error Rate (%)

Number of Features

Distances

Number of Features
Edge Maps

Improvements at Low Error Rates

Concavities  Distances  

Edge Maps  Combination
Concavities

Distances
Edge Maps

Combination of the three Ensembles
Some Errors

Summary

• Two-level MOGA
  – 1st → Feature selection to promote diversity among the classifiers.
  – 2nd → Combines those classifiers to yield powerful ensembles.

• Compared to Optiz’s Method, our’s:
  – Generate multiple ensembles.
  – Smaller ensembles.
  – Optiz uses the size of the population (20 classifiers).
Future Work

- Combining heterogeneous ensembles
  - Investigate the use of unsupervised learning in the context of the supervised learning [Ho’02].