
L. S. Oliveira, R. Sabourin, F. Bortolozzi, and C.Y. Suen

École de Technologie Supérieure, Montreal, Canada.
Centre for Pattern Recognition and Machine Intelligence, Montreal, Canada.
Pontifícia Universidade Católica do Paraná, Curitiba, Brazil.
Introduction

• To identify the best subset of features to represent a pattern from a larger set often mutually redundant of even irrelevant features.
  – Minimize the error rate of the classifier.
  – Minimize the number of features.
• Interdependence: two or more features convey important information.
• Classical methods:
  – Features evaluated on their individual merits.
  – Ignore interactions between features.
Introduction

- 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

Pareto-optimal front
Multi-Objective GA

• 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

• 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

- Proposed by Srinivas & Deb 95.
- Ranking by fronts.
- It converges close to the Pareto-optimal front.
Flow Chart of the Methodology

1. Generate Initial Population
2. Poll of Candidates
   - Apply NSGA and genetic operators
   - New Pool of Candidates
3. Neural Network and Sensitivity Analysis
4. Rank according to fitness values
5. Pareto-optimal front
   - Pareto-optimal front with its respective Validation Curve
6. Validation
7. Training Solution Indicated by the Validation Curve
8. Optimized feature vector
Methodology

• **NSGA**
  – Bit representation, one-point crossover, bit-flip mutation, and elitism.

• **Fitness evaluation:**
  – Number of selected features.
  – Error rate of the classifier.
Methodology

• Sensitivity analysis.
  – 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 Classifier

- 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) – 99.13% (zero rej. level).
- Feature set:
  - Concavities and contour (132 components).
- More details see PAMI vol. 24, n. 11, 2002.
Experiments

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

- Pareto-based approach.
  - It converges close to the Pareto-optimal front.
  - Importance of validating the Pareto-optimal front.
Results

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

<table>
<thead>
<tr>
<th>Zero-rejection level</th>
<th>Original Classifier</th>
<th>Optimized Classifier</th>
</tr>
</thead>
<tbody>
<tr>
<td>Features</td>
<td>R.R.(%)</td>
<td>Features</td>
</tr>
<tr>
<td>132</td>
<td>99.13</td>
<td>100</td>
</tr>
</tbody>
</table>

Error rate fixed at 0.5%

<table>
<thead>
<tr>
<th>Original Classifier</th>
<th>Optimized Classifier</th>
</tr>
</thead>
<tbody>
<tr>
<td>Features</td>
<td>R.R.(%)</td>
</tr>
<tr>
<td>132</td>
<td>98.50</td>
</tr>
<tr>
<td>100</td>
<td>98.54</td>
</tr>
</tbody>
</table>
Conclusion

• Methodology for feature selection.
  – Modified wrapper.
    • Sensitivity analysis with neural networks.
  – Validation set to point out the best solution of the Pareto-optimal front.

• Advantages of multi-objective GA.
  – It avoids dealing with problems such as weighting and scaling objectives.
  – Provides a set of potential solutions.
Conclusion

• Reduced feature set:
  – 25% less features with the same performance.

• Future works:
  – Feature selection for ensembles.