From Mating Pool Distributions to Model Overfitting

Claudio F. Lima\textsuperscript{1}  Fernando G. Lobo\textsuperscript{1}  Martin Pelikan\textsuperscript{2}

\textsuperscript{1}Department of Electronics and Computer Science Engineering  
University of Algarve, Portugal

\textsuperscript{2}Missouri Estimation of Distribution Algorithm Laboratory  
University of Missouri at St. Louis, USA

GECCO 2008, Atlanta, USA
Probabilistic Modeling in EDAs

The core
- EDA is defined by its **model complexity** (expressiveness).

The spectrum
- Simpler EDAs only learn the **parameters** of the model.
- More powerful EDAs learn both **structure and parameters**.

Bayesian EDAs
- Solve a broad class of problems **efficiently**.
- But oftentimes models do not exactly reflect the **underlying problem structure**.
Motivation

Importance of model accuracy

- Model-based efficiency enhancement techniques.
  - Substructural local search.
  - Fitness estimation (surrogate fitness function).
- Offline interpretation of probabilistic models.
  - Bias structural learning in EDAs (Mark Hauschild’s talk).
  - Design structure-aware variation operators.

Empirical findings with BOA: Tournament vs. truncation

- Population size requirements.
- Model quality difference.
Bayesian Optimization Algorithm (BOA)

**BOA**
- EDA which uses Bayesian networks (BNs).
  - Model selected solutions to generate new ones.
- Similar algorithms:
  - Estimation of Bayesian networks algorithm (EBNA).
  - Learning factorized distribution algorithm (LFDA).

**Learning BNs**
- Decision trees express conditional probabilities.
- Greedy algorithm for learning.
- Scoring metrics considered: K2 and BIC.
Experimental Setup

Test problem
- To investigate MSA we use a problem of known structure.
- Clear identification of:
  - Necessary dependencies $\rightarrow$ successful tractability.
  - Unnecessary dependencies $\rightarrow$ poor model interpretability.

$m - k$ trap problem
- Consists of $m$ concatenated $k$-bit trap functions.
- For $k \geq 3$, the model should be able to maintain $k$–order statistics, otherwise $\rightarrow$ exponential scalability.
- $k = 5$ used in experiments.
Ideal Model for the $m - k$ trap problem

Joint distribution in BNs

$$p(X_a, X_b, X_c, X_d) = p(X_a)p(X_b|X_a)p(X_c|X_a, X_b)p(X_d|X_a, X_b, X_c)$$

- **Clique** between variables of the same trap function.
- While between different traps there are no dependencies.
Measuring Model Structural Accuracy

**Definition 1**

The model structural accuracy (MSA) is defined as the ratio of correct edges over the total number of edges in the BN.

**Definition 2**

An edge is correct if it connects two variables that are linked according to the objective function definition.

**Definition 3**

Model overfitting is defined as the inclusion of incorrect (or unnecessary) edges to the BN.
Selection Methods

Selection methods considered

- **Tournament selection.** Randomly pick $s$ individuals and choose the best. Repeat $n$ times.

- **Truncation selection.** Choose the best $\tau$ proportion of the population.

Equivalent selection intensity

<table>
<thead>
<tr>
<th>Selection intensity, $I$</th>
<th>0.56</th>
<th>0.84</th>
<th>1.03</th>
<th>1.16</th>
<th>1.35</th>
<th>1.54</th>
<th>1.87</th>
</tr>
</thead>
<tbody>
<tr>
<td>Tournament size, $s$</td>
<td>2</td>
<td>3</td>
<td>4</td>
<td>5</td>
<td>7</td>
<td>10</td>
<td>20</td>
</tr>
<tr>
<td>Truncation threshold, $\tau$ (%)</td>
<td>66</td>
<td>47</td>
<td>36</td>
<td>30</td>
<td>22</td>
<td>15</td>
<td>8</td>
</tr>
</tbody>
</table>
Selection-Metric Comparison

Observations

1. Truncation performs better than tournament selection.
2. K2 metric performs better than BIC metric.
A Simple Analysis of Tournament Selection

Probability of an individual with rank \( i \) to win a tournament

\[
p_i = \prod_{j=1}^{s-1} \frac{i-j}{n-j}, \quad \text{for } s \geq 2. \tag{1}
\]

Number of copies in the mating pool

\[
c_i = s \ p_i. \tag{2}
\]

Which can be approximated by a power distribution with p.d.f.

\[
f(x) = \alpha \ x^{\alpha-1}, \quad 0 < x \leq 1, \ \alpha = s, \ x = i/n. \tag{3}
\]
A Simple Analysis of Truncation Selection

Number of copies in the mating pool

\[ c_i = \begin{cases} 
0, & \text{if } i < n \left( 1 - \frac{\tau}{100} \right) \\
1, & \text{otherwise.} 
\end{cases} \] (4)
**Introduction**

Selection as a Source of Overfitting

Improving Model Accuracy

**Influence of Selection in Model Accuracy**

Selection: The Mating-Pool Distribution Generator

---

**Distribution of the Expected Number of Copies**

![Graph showing the distribution of the expected number of copies for different ranks and selection parameters.](image)

**Tournament and truncation selection differ in:**

1. **Window size.**
2. **Distribution shape.**

---

Lima, Lobo, & Pelikan

From Mating Pool Distributions to Model Overfitting
Mating Pool Distribution: An Example

Function

\[ f_{mk-trap} = f_1(X_1X_2X_3) + f_2(X_4X_5X_6) + f_3(X_7X_8X_9) \]

Example

<table>
<thead>
<tr>
<th>Rank</th>
<th>Individual</th>
<th>Copies</th>
</tr>
</thead>
<tbody>
<tr>
<td>8</td>
<td>111 100 000</td>
<td>1 1 4 8</td>
</tr>
<tr>
<td>7</td>
<td>111 111 101</td>
<td>1 1 3 4</td>
</tr>
<tr>
<td>6</td>
<td>111 000 110</td>
<td>1 1 2 2</td>
</tr>
<tr>
<td>5</td>
<td>000 001 000</td>
<td>1 1 1 1</td>
</tr>
<tr>
<td>4</td>
<td>111 010 011</td>
<td>0 0 0 0</td>
</tr>
<tr>
<td>3</td>
<td>000 101 001</td>
<td>0 0 0 0</td>
</tr>
<tr>
<td>2</td>
<td>000 110 101</td>
<td>0 0 0 0</td>
</tr>
<tr>
<td>1</td>
<td>101 000 101</td>
<td>0 0 0 0</td>
</tr>
</tbody>
</table>

Mating pool distributions:

- Uniform: \( X_6X_7 \) freq.
- Linear: \( X_6X_7 \) freq.
- Exponential: \( X_6X_7 \) freq.

<table>
<thead>
<tr>
<th>Rank</th>
<th>Individual</th>
<th>Copies</th>
</tr>
</thead>
<tbody>
<tr>
<td>8</td>
<td>111 100 000</td>
<td>0.25 0.4 0.53</td>
</tr>
<tr>
<td>7</td>
<td>111 111 101</td>
<td>0.25 0.2 0.13</td>
</tr>
<tr>
<td>6</td>
<td>111 000 110</td>
<td>0.25 0.1 0.07</td>
</tr>
<tr>
<td>5</td>
<td>000 001 000</td>
<td>0.25 0.3 0.27</td>
</tr>
</tbody>
</table>
Mating Pool w/ Non-Uniform Distribution

What’s good about it?
- Detection of existent patterns is achieved with less data.
  - Tournament selection requires smaller population sizes than truncation.

What’s bad about it?
- Irrelevant features can also be detected as patterns.
  - Excessive complexity → model overfitting.

So what?
- Recognize that we are learning from imbalanced data sets.
- Change scoring metric accordingly!
Improving Model Accuracy for Tournament

Agenda

1. Compute how the power-law-based frequencies differ from uniform ones in the mating pool.
2. Use computed frequency deviation to estimate scoring metric gain due to overfitting.
3. Change the complexity penalty to counterbalance the misleading metric gain.
Deviation of power-law-based frequencies

Deviation is linear w.r.t. tournament size

- Expected market share of top-ranked individuals after selection is given by the (right-side area of the) c.d.f. of the power distribution.

\[ F(x) = x^s, \quad 0 < x < 1, \quad s \geq 2. \]  

- In the worst case, market share grows linearly with tournament size.
- Thus, misleading frequencies deviate linearly from the “true” ones (uniform distribution).
- More details in the paper.
Consider the decision of adding an edge from $X_2$ to $X_1$

- Metric gain must be calculated:
  \[ G_{\text{metric}} = \text{ScoreAfter} - \text{ScoreBefore} - \text{ComplexityPenalty} \]

- Uniform mating pool reveals
  \[ p_{00} = p_{01} = p_{10} = p_{11} = 0.25 \rightarrow G_{\text{metric}} < 0. \]

- Simulate tournament mating pool by assigning
  \begin{align*}
  p_{00} &\approx 0.25 + \Delta(s - 1), \\
  p_{01} &\approx 0.25 - \Delta(s - 1), \\
  p_{10} &\approx 0.25 - \Delta(s - 1), \\
  p_{11} &\approx 0.25 + \Delta(s - 1).
  \end{align*}

- Where $\Delta$ is related to the proportion of top-individuals
  which lead to overfitting (more info in the paper).
Approximated metric gain

\[ G' \approx 2 \left( 0.25 - \Delta(s - 1) \right) \log_2 \left( 0.25 - \Delta(s - 1) \right) + 1. \]  (6)
Comparison of Different Penalty Corrections

Complexity penalty for the K2 metric

\[ p(B) = 2^{-0.5c_s \log_2(n) \sum_{i=1}^{\ell} |L_i|} \rightarrow c_s = \{ 1, \sqrt{s}, s, and s \log_2(s) \} \]
Results for the $s$–penalty ($c_s = s$)

Remark
- Model structural accuracy is superior to 99%!
Results for Truncation with the Standard Penalty

Remark
- Good results, but not as good as for tournament selection with the $s$–penalty.
Summary & Conclusions

- Selection addressed as a source of overfitting.
- Influence of selection method in BN learning:
  - Truncation selection → uniform mating pool is more appropriate for learning (standard scoring metrics).
  - Tournament selection → non-uniform mating pool leads to model overfitting.
- Scoring metric adapted to the power distribution of the mating pool generated by tournament selection.
  - Model accuracy has been considerably improved!
- Improved model interpretability to assist efficiency enhancement techniques and human researchers.
From Mating Pool Distributions to Model Overfitting

Claudio F. Lima¹  Fernando G. Lobo¹  Martin Pelikan²

¹Department of Electronics and Computer Science Engineering
University of Algarve, Portugal

²Missouri Estimation of Distribution Algorithm Laboratory
University of Missouri at St. Louis, USA

GECCO 2008, Atlanta, USA
Scoring Metrics

(Bayesian-Dirichlet) K2 metric

\[
K2(B) = p(B) \prod_{i=1}^{\ell} \prod_{l \in L_i} \frac{1}{\Gamma(m_i(l) + 2)} \prod_{x_i} \Gamma(m_i(x_i, l) + 1),
\]

where \( p(B) = 2^{-0.5 \log_2(n) \sum_{i=1}^{\ell} |L_i|}. \)

Bayesian information criteria (BIC) metric

\[
BIC(B) = \sum_{i=1}^{\ell} \left( \sum_{l \in L_i} \sum_{x_i} \left( m_i(x_i, l) \log_2 \frac{m_i(x_i, l)}{m_i(l)} \right) - |L_i| \frac{\log_2(n)}{2} \right).
\]