Parameter Cross-validation and Early-stopping in Univariate Marginal Distribution Algorithm

Hao Wu
School of Computer Science
The University of Manchester
Manchester, M13 9PL, UK
Tel: +44 161 2756205
wuh@cs.man.ac.uk

Jonathan L. Shapiro
School of Computer Science
The University of Manchester
Manchester, M13 9PL, UK
Tel: +44 161 2756253
jls@cs.man.ac.uk

ABSTRACT
In this paper, a cross-validation and early-stopping algorithm is devised for parameter updating in the Univariate Marginal Distribution Algorithm (UMDA) to reduce overftting. Our hypothesis is that the well-known problem of diversity loss in UMDA is a consequence of overfitting during the parameter estimation step at each generation. It is tested by experiments on two different optimization problems.

Categories and Subject Descriptors
I.2.8 [**Artificial Intelligence**]: Problem Solving, Control Methods, and Search

General Terms
Algorithms

Keywords
Cross-validation, Early-stopping, Overfitting, UMDA

1. INTRODUCTION
Estimation of Distribution Algorithms (EDAs) [1] are a class of evolutionary algorithms that use machine learning techniques to solve optimization problems. One known problem of EDAs is that they may suffer from a form of "premature convergence", in which diversity is lost from the probability model and is not restored. For parameter learning, most of EDAs choose the parameters which maximize the likelihood of the data. It has recently been shown that all EDAs which use maximum likelihood for parameter estimations suffer from this premature convergence effect[2]. Maximum likelihood always results in expected diversity loss when going from one generation to next.

In this paper, we propose to use cross-validation in parameter learning of EDAs to reduce premature convergence. Cross-validation is a method to avoid overfitting in machine learning. Our hypothesis is that premature convergence in EDAs is caused by overfitting iterated many times. Recent research has shown that overfitting does occur during the structure learning in EDAs with small population sizes, and this is not fully removed by regularization methods such as the Bayesian Information Criterion (BIC) [3]. It was further shown that overfitting correlates with algorithm performance, and that performance can be improved by the use of cross-validation and early stopping. Overfitting happens in structure learning, but it has not been investigated whether it occurs in parameter learning? Here, we use cross-validation and early-stopping for parameter learning one of classical EDAs, Univariate Marginal Distribution Algorithm (UMDA). For that the model only requires parameter learning.

GECCO'07, July 7–11, 2007, London, England, United Kingdom
ACM 978-1-59593-697-4/07/0007.

2. A CROSS-VALIDATION ALGORITHM FOR UMDA
Assuming that the search space consists of L binary variables, x_i, where $i = 1,\ldots L$. The joint probability of UMDA model is

$$P(\{x_1,...,x_L\}) = \prod_{i=1}^{L} P(x_i) \quad (1)$$

The L parameters of the model are $P(x_i = 1) \equiv \gamma_i$.

In standard UMDA: $\gamma_i \leftarrow \frac{1}{N}\sum_{X\in D} x_i$ (2)

In our algorithm: $\gamma_i \leftarrow \gamma_i + \alpha(\frac{1}{N}\sum_{X\in D} x_i - \gamma_i)$ (3)

Where D denotes the dataset and $\alpha \in [0,1]$. An independent dataset called validation set is used to optimize α, whose size is equal to the training dataset for modeling.

3. EXPERIMENTS
Experiments on needle-in-a-haystack problem and basin-with-a-barrier problem show that overfitting exists in parameter estimation of UMDA; the cross-validation and early-stopping method could slow down the premature convergence in parameter estimation of UMDA and improve algorithm's performance.

4. REFERENCES
[1] Pedro Larranaga, Jose A. Lozano. *Estimation of Distribution Algorithms*. University of the Basque Country, Spain, 2002.

[2] J. L Shapiro, *Diversity loss in general estimation of distribution algorithms*. In Parallel Problem Solving from Nature, volume 4193 of LNCS, pages 92-101. 2006.

[3] Hao Wu and J.L. Shapiro, *Does overfitting Affect performance in Estimation of Distribution Algorithms*. GECCO'06, Seattle, USA 2006