# Does Overfitting Affect Performance in Estimation of Distribution Algorithms

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

Estimation of Distribution Algorithms (EDAs) are a class of evolutionary algorithms that use machine learning techniques to solve optimization problems. Machine learning is used to learn probabilistic models of the selected population. This model is then used to generate next population via sampling. An important phenomenon in machine learning from data is called *overfitting*. This occurs when the model is overly adapted to the specifics of the training data so well that even noise is encoded. The purpose of this paper is to investigate whether overfitting happens in EDAs, and to discover its consequences. What is found is: overfitting does occur in EDAs; overfitting correlates to EDAs performance; reduction of overfitting using early stopping can improve EDAs performance.

## Categories and Subject Descriptors

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

## General Terms

Algorithms

## Keywords

Overfitting, Estimation of Distribution Algorithms, Bayesian Optimization Algorithm, Random 3-SAT

## 1. INTRODUCTION

Estimation of Distribution Algorithms (EDAs) [1] are a class of evolutionary algorithms that use machine learning techniques to solve optimization problems. They generally build probabilistic graphical model based on good solutions found so far and use the constructed models to guide the further search.

One of the most important phenomena in machine learning from data is ***overfitting***. In this, the learning algorithm adapts so well to the given data, that noise or particularities of the specific sample are also encoded by the learned model. It results in reduced performance when the task is the generalization to unseen data, as well as producing an overly complex model which may consume unnecessary learning time and computational resources.

Overfitting can be observed by cross-validating against a validation set which is independent of the set used for learning. In general, learning involves reducing some loss function. During learning, the loss function evaluated on the training data is reduced as the model fits the training data better and better. However, when the loss function is computed on validation, it often starts to decrease, but after a time starts to increase again, as shown in Figure 1. When the loss function on the validation data starts to increases, it indicates that the learner is fitting the specifics of the training data only, and is overfitting.

Standard methods for building probability models from data use regularization methods to avoid overfitting. This is done by adding a penalty term to the loss function, such as the Bayesian information criterion (BIC), which penalizes overly complex models, or by integrating over the parameters. Regularization does not fully avoid overfitting in small samples, as Section 3 shows. Since EDAs use machine learning techniques to construct model, is overfitting also important in EDAs? As far as we know, this question has not been investigated. Since the goal of EDAs is optimization not generalization to unseen data, perhaps it is understandable that it has not been considered. On the other hand, recent work has shown that EDAs do need some regularization to avoid being overly influenced by noise [2]. The goal of this paper is to determine whether overfitting is important in EDAs.

## 2. METHODOLOGY

### 2.1 Cross-validation

In machine learning, a method called *Cross-validation* [3] can be used to observe overfitting. First, the data is divided into two sets: a training set and a validation set. The training set is used to construct the model, and the validation set is used to estimate the true performance. The performance of model is calculated using both sets at each training cycle. As the model becomes more complex, the performance of the model as measured on the training set increases. However, model will not fit the validation data any better past a certain model complexity, at which point overfitting occurs, as shown in Figure 1.

### 2.2 Bayesian Optimization Algorithm (BOA)

The EDA investigated herein is called the Bayesian Optimization Algorithm (BOA) [4]. This algorithm is appropriate for discrete domains. It uses Bayesian networks (BN) to model selected data.

### 2.3 Random 3-SAT Problem

The problem used to test overfitting in EDAs is Random 3-SAT. Propositional satisfiability (SAT) is the problem of deciding whether there is a configuration for the Boolean variables in a

Copyright is held by the author/owner(s).
*GECCO'06*, July 8–12, 2006, Seattle, Washington, USA.
ACM 1-59593-186-4/06/0007.

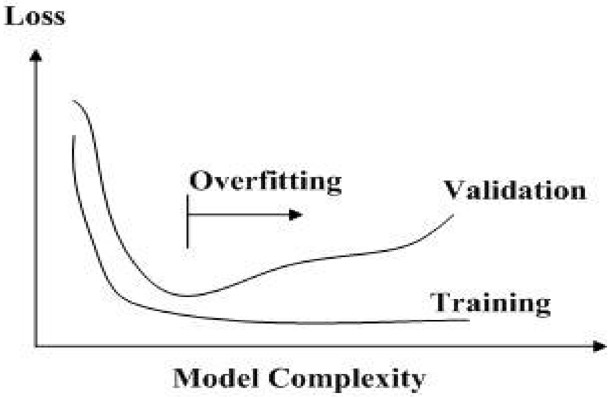

**Figure 1. As model complexity increases, performance on the data used to build the model (training data) improves. However, performance on an independent set (validation data) improves up to a point, then starts to get worse. This is called overfitting.**

propositional formula that makes the formula true. SAT problems are normally considered in Conjunctive Normal Form (CNF): a formula is in CNF if it is a conjunction of clauses, where a clause is a disjunction of literals, and a literal is a negated or un-negated variable. A problem in random *k*-SAT consists of *m* clauses, each of which has *k* literals chosen uniformly from the *n* possible variables and the n possible negated variables. When *k* equals to 3, the problem is called random 3-SAT. It is known that 3-SAT has a phase transition between satisfiability and unsatisfiability when the ratio of the number of clauses over the number of variables ***m/n*** is around 4.25.

## 3. OVERFITTING IN EDAS

The following results show that overfitting does happen in EDAs when small samples are used, even when a standard regularization is performed to avoid it. The problem size *n* is set to 15. To make these SAT problems hard, the ratio ***m/n*** is set to 4.25. A comparatively small population size is used, *N=100*. Greedy search with edge addition only is used to learn the model. At each learning step, an edge is added if its addition increases the score, which is the sum of the log-likelihood and the BIC regularizer to penalize for an overly complex model. At each generation, we sampled the present model twice to generate a training dataset and a validation set, and calculated the likelihood $P(Data \mid Model)$ for both. The likelihood shows how likely it was for the model to have generated the data; we assume that when the likelihood of the validation data decreases, the model is overfitting the training data. Results on 50 random 3-SAT problems run for 10 generations each show that overfitting happens with high intensity from the generation 1 to the generation 7. The intensity of overfitting decreased generation by generation and nearly disappeared at generation 8. Figure 2 is the result of Generation 3.

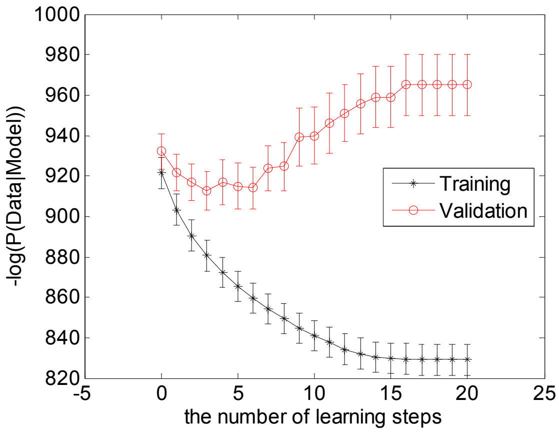

**Figure 2. Overfitting in generation 3 of a 15 variables random 3-SAT problem, averaged over 50 runs.**

## 4. CORRELATION BETWEEN OVERFITTING & PERFORMANCE

As Figure 1 shows, overfitting increases the gap between the validation and training curve. Does this influence correlate to the performance in EDAs? We performed BOA to solve random 3-SAT problems multiple times, and made linear regression with gap between training and validation sets as independent variable, and the difference in fitness between two subsequent generations as the dependent variable. We calculated the confidence interval. The slope of the regression line is negative with 98% confidence interval in 10 different problems. This shows the tendency towards greater improvement in fitness between generations when there is a smaller gap in the log-likelihoods of the two data sets.

## 5. REDUCING OVERFITTING TO IMPROVE EDAS PERFORMANCE

In this section, one of most widely used methods, *early stopping*, is chosen to reduce overfitting. It uses the training and validation set to watch overfitting. When overfitting is detected by the rise on validation set, the learning is immediately terminated. The statistic results below show that the algorithm performance did get improvement when overfitting is reduced in a simple way. The problem size *n* is varied by 10, 15 and 20. We used both BOA without overfitting control and BOA with reducing overfitting to solve 200 randomly generated 3-SAT problems.

**Table 1. Results of BOA on 200 random 3-SAT problems**

| Problem Size | 10 | 15 | 20 |
|---|---|---|---|
| Number solved without early stopping | 190 | 156 | 118 |
| Number solved with early stopping | 192 | 175 | 135 |

## 6. CONCLUSION

We have shown that overfitting happens in EDAs and has an important effect on performance. When simple early stopping is used, improvement is small, perhaps suggesting the need for other methods to be developed and other problems to be considered.

## 7. REFERENCES

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

[2] J. L Shapiro, *Drift and scaling in estimation of distribution algorithms*. Evolutionary Computation, 13(1), p99-123, 2005.

[3] Tom M. Mitchell. *Machine Learning*. London; New York: WCB/McGraw-Hill, 1997.

[4] Martin Pelikan. *Bayesian Optimization Algorithm: From Single Level to Hierarchy*. Ph.D. Thesis, University of Illinois at Urbana-Champaign, Urbana, Illinois, 2002.