Performance Evaluation of Mutation Operators in Estimation of Distribution Algorithms: A Study in Bayesian Structure Learning

Song Ko[1], Dae-Won Kim[1], and Bo-Yeong Kang[2⋆]

[1]School of Computer Science and Engineering, Chung-Ang University, Seoul, Korea
[2]School of Mechanical Engineering, Kyungpook National University, Daegu, Korea
kby09@knu.ac.kr

Abstract. Estimation of distribution algorithms (EDAs) have recently shown an efficient optimization performance in various problem domains. However, few reports have examined the advantages of using a mutation operator in EDAs for learning Bayesian networks. In this paper, we extended three representative EDAs based on mutation operators and evaluated their performances in Bayesian structure learning. The experimental results show that the mutation-adopted EDAs give markedly better performance than the conventional algorithms. The PBIL+BT algorithm gives the most accurate structure of Bayesian networks.

Key words: Bayesian networks, Estimation of distribution algorithms

1 Introduction

Estimation of distribution algorithms (EDAs) are a novel paradigm for evolutionary computation that were developed as a natural alternative to genetic algorithms (GAs). The advantage of EDAs over GAs are the absence of variation parameters and the expressiveness of the probabilistic model that guides the search process [1]. With respect to Bayesian structure learning, Blanco et al. [2] first used EDAs to infer the structure of Bayesian networks, and showed that competitive results are achieved by EDAs over GAs. The heuristic scheme of EDAs has been proved to be effective and efficient in Bayesian structure learning [2]. In [3–5], it is reported that the incorporation of mutation operator into EDAs can increase the diversity of genetic information in the generated population; the use of bit-flip mutation was shown to be effective in Four-peaks problems and MAXSAT problems.

Users of EDAs are often left wondering to what extent the performance of EDAs can be improved by exploiting mutations and which EDA algorithm is the most improved by mutations for learning Bayesian networks. In this paper, we evaluate the performance of EDAs with mutation to provide answers to these questions. To our knowledge, this is the first such comparison of EDAs with

⋆ corresponding author

Algorithm 1 Estimation of Distribution Algorithm

Generate N individuals randomly for the first generation ($t \leftarrow 0$

ividuals according to their ranks

Estimate a probabilistic model for M individuals

Generate N new individuals by sampling the probabilistic model $t \leftarrow t + 1$

until Stopping criteria is met

mutation for Bayesian structure learning. We consider three well-known EDAs: PBIL, MIMIC, and BOA. We compare their performance with those of their mutation-adopted versions. As mutation operators, the bit-flip in [3–5], a newly proposed transpose mutation, and their combinations are employed. These comparisons are made by applying the algorithms to two public data sets.

2 Related Work

Algorithm 1 shows the procedural steps of EDAs, which repeat a set of genetic operations (e.g., sampling, evaluation, selection, and estimation) until a stopping criterion (e.g., the number of generations, a running time, and a specific fitness value) is satisfied. GAs rely on crossover and mutation operators to generate offspring from two or more individuals, whereas EDAs use a probabilistic model estimated from the selected individuals. The new population of individuals is sampled from a probabilistic model, which is estimated from representative individuals selected from the previous generation. The interdependencies between variables in each individual are explicitly expressed by the joint probability distribution of the individuals selected at each generation.

EDAs can be divided according to the complexity of the probabilistic models used to capture the dependencies between variables: univariate, bivariate, and multivariate approaches. In this study, three widely-used EDAs are used to represent these approaches. In PBIL [6], the joint probability distribution of variables is constructed by a product of univariate marginal probability distribution. MIMIC [7] searches for the best permutation between the variables using a chain structure to capture the pairwise dependencies between variables. BOA [8] factorizes the joint probability distribution using statistics of order greater than two. To represent multiple relations between variables, the Bayesian Dirichlet equivalent metric is used to measure the goodness of each candidate dependency.

In [2], Blanco et al. used the PBIL to infer the structure of Bayesian networks. The structure learning of Bayesian networks is a NP-Hard problem since the number of structures grows exponentially with the number of variables. Because of the huge cardinality of the search spaces, the heuristic approach of EDAs has been shown to be effective and efficient. In [9], the MIMIC was used to obtain the optimal ordering of variables for Bayesian networks. Compared to GAs, the PBIL and MIMIC provided the more similar Bayesian networks with respect to the original networks.

3 EDAs with Mutations in Bayesian Learning

In EDAs, we construct a probabilistic model of feasible solutions in the search space, and use this model to guide the evolutionary search for the optimal solution. To achieve this purpose, it is important to keep the diversity of individuals in the generated populations so that it can avoid premature convergence to a sub-optimal solution. Recent studies have proposed new EDAs that use bit-flip mutations to conventional EDAs to increase the population diversity [3–5]. In this study, we show the effectiveness of mutation-adopted EDAs in identifying an optimal Bayesian network structure. To this end, we propose a new mutation operator, called transpose mutation, designed for Bayesian structure learning.

To introduce new randomness that was not considered in the probable solutions, we present a transpose mutation that generates offsprings by inverting the arc direction in the individuals, which can inherit the information of solutions to guide the search over probable solutions as well as explore new states for offsprings by changing the direction of arcs. For representation of the k_{th} individual (C^k) in Bayesian networks, the $n \times n$ binary connectivity matrix is used; n is the number of variables. Each element C_{ij} in C^k is defined as 1 if a variable i is a parent of a variable j, 0 otherwise:

$$C^k = \begin{pmatrix} 0 & C_{12} & \cdots & C_{1n} \\ \vdots & \vdots & \ddots & \vdots \\ C_{n1} & C_{n2} & \cdots & 0 \end{pmatrix}$$

Using the matrix representation, we build a probability matrix (P) that estimates the probability distribution of dependencies (arcs) between variables (nodes). Given a population $C = \{C^1, C^2, \ldots, C^N\}$, P is constructed using $T = \{T^1, T^2, \ldots, T^M\}$ ($T \subseteq C$), which are selected according to their fitness ranks. P_{ij} ($0 \leq P_{ij} \leq 1$) represents how often an arc occurs in the selected individuals.

$$P^k = \begin{pmatrix} 0 & P_{12} & \cdots & P_{1n} \\ P_{21} & 0 & \cdots & P_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ P_{n1} & P_{n2} & \cdots & 0 \end{pmatrix}$$

where $P_{ij} = 1/M(T^1_{ij} + T^2_{ij} + \ldots + T^M_{ij})$. The offspring O is generated by the probability model and mutations until the number of offsprings becomes N. To generate an arc (O_{ij}) of an offspring O, it compares two probability values, i.e. P_{ij} and $rand[0,1]$:

$$O_{ij} = \begin{cases} 1 & \text{if } P_{ij} > rand[0,1] \\ 0 & \text{otherwise.} \end{cases}$$

Then, in case of the bit-flip mutation, the directions of arcs are flipped (switching between 0 to 1) according to a mutation rate (γ):

$$O_{ij} = \begin{cases} (O_{ij}+1)\%2 & \text{if } \gamma > rand[0,1] \\ O_{ij} & \text{otherwise.} \end{cases}$$

In case of the transpose mutation, it inverts the direction of arcs by replacing O_{ij} with $[O_{ij}]^T$ according to the value of γ:

$$O = \begin{cases} O^T & \text{if } \gamma > rand[0,1] \\ O & \text{otherwise.} \end{cases}$$

The transpose mutation explores the conditional dependencies among variables in reverse order; a reversal operator has been widely used for the traveling salesman problem due to the possibility for escaping local optima [10]. Moreover, for offsprings with invalid directed acyclic graphs, we repaired it by removing a violated arc at random.

4 Experimental Results

To test the mutation-adopted EDAs, we applied three EDAs as well as their mutation-adopted versions to two widely used data and compared the performances of the algorithms; the conventional EDAs were PBIL, MIMIC, and BOA; the bit-flip mutation versions were PBIL+B, MIMIC+B, and BOA+B; the transpose mutation versions were PBIL+T, MIMIC+T, and BOA+T; both mutation adopted versions were PBIL+BT, MIMIC+BT, and BOA+BT.

The data sets employed were the Asia and Diabetes data; they have been widely used for comparative purposes in Bayesian structure learning. The Asia [11] data set contains the relevant variables and relationships for medical knowledge related to the shortness of breath (dyspnoea). The network contains eight variables and eight arcs. The Diabetes [12] data set is a diagnostic network for predicting the signs of diabetes. The Diabetes network contains nine variables and 11 arcs. Each network has been used to generate sample cases, each of which contains 10,000 instances. In these experiments, the well-known BDe score for Bayesian networks was used as the fitness function. The additional parameters were set as: the number of generations was set to 400; six mutation rates were used ($\gamma = 0, 0.01, 0.05, 0.1, 0.2, 0.3$); the population size was set to 50; each experiment was run 30 times. We listed the F-measure achieved by each EDA for the two data sets in Tables 1 and 2. We assessed the F-measure by comparing the network structures inferred by EDAs with respect to the original network.

For the Asia data set (Tables 1), the bit-flip and transpose mutation-adopted EDAs showed better performance than their conventional versions ($\gamma = 0$) in terms of F-measure. Of the two mutations versions, the EDAs+T showed slightly better improvement than EDAs+B. Notably, the both mutation-adopted EDAs gave markedly better performance than their conventional counterparts. Of particular note are the F-measure results for the PBIL+BT ($\gamma = 0.1$), for which the PBIL+B and PBIL+T gave improved F-measures of 46.6% and 67.7%, respectively, compared to the conventional PBIL (37.7%), but the PBIL+BT exhibited 87.6%.

For the Diabetes data set (Tables 2), the EDAs+B and EDAs+T provided better performance than their conventional algorithms ($\gamma = 0$). Meanwhile, the

Table 1. Comparison of F-measure(%) achieved by EDAs for the Asia data.

Mutation	Bit-flip			Transpose			Bit-flip+Transpose		
(γ)	PBIL	MIMIC	BOA	PBIL	MIMIC	BOA	PBIL	MIMIC	BOA
0	37.7	27.2	25.3	37.7	27.2	25.3	37.7	27.2	25.3
0.01	41.7	28.0	26.9	46.8	27.2	26.1	49.9	33.2	30.5
0.05	44.2	31.2	24.1	59.1	27.8	27.7	65.0	31.2	25.2
0.1	46.6	40.4	28.3	67.7	34.3	37.4	87.6	37.8	42.6
0.2	48.8	33.9	28.1	63.7	32.7	31.2	82.2	36.3	49.1
0.3	65.7	24.8	28.3	66.5	36.9	31.4	84.4	41.4	52.8

Table 2. Comparison of F-measure (%) achieved by EDAs for the Diabetes data.

Mutation	Bit-flip			Transpose			Bit-flip+Transpose		
(γ)	PBIL	MIMIC	BOA	PBIL	MIMIC	BOA	PBIL	MIMIC	BOA
0	36.6	31.2	25.5	36.6	31.2	25.5	36.6	31.2	25.5
0.01	48.3	37.8	27.6	54.5	33.5	24.6	58.5	35.8	34.0
0.05	47.0	39.4	27.2	64.3	29.6	30.0	76.1	33.0	30.8
0.1	48.9	40.1	24.6	64.6	35.2	34.7	86.3	38.3	39.7
0.2	66.3	43.4	33.3	66.3	39.3	44.4	77.8	45.3	41.6
0.3	73.5	40.3	36.4	64.4	31.9	48.2	75.2	43.3	50.8

F-measure results of the EDAs+BT were superior to those of their counterparts. Especially, the PBIL+BT (γ=0.1) showed significantly better performance than PBIL+B and PBIL+T, giving F-measure of 86.3%.

In Table 3, we compared the proportion of arcs in the inferred networks by EDAs with mutations for the Diabetes data ($\gamma = 0.1$) in terms of the proportions of correct arcs and reverse arcs inferred by each algorithm, respectively, and the proportions of additionally inferred arcs that are non-existent in the original network. It is evident that the PBIL was superior to MIMIC and BOA. In particular, PBIL+BT gave markedly better performance than the other methods; the proportions of correct arcs were significantly increased and the proportions of reverse and additional arcs were dramatically decreased, yielding to the higher values of F-measures. The results of the comparison calculations indicate that the PBIL+BT gave the greatest enhancement of performance in Bayesian structure learning. These results thus highlight the effectiveness and potential utility of the mutation-adopted EDAs.

5 Conclusion

Compared to the conventional EDAs, the mutation-adopted EDAs showed better performance for learning Bayesian networks. The present evaluation has shown that PBIL+BT gave the greatest enhancement of learning performance. However, several issues require further investigation. Although the transpose mutation is considered powerful, further work investigating the theoretical foundation for its effectiveness in Bayesian structure learning is needed. Secondly, despite

Table 3. The proportion (%) of arcs inferred by EDAs for the Diabetes data

Arc type	Bit-flip			Transpose			Bit-flip+Transpose		
	PBIL	MIMIC	BOA	PBIL	MIMIC	BOA	PBIL	MIMIC	BOA
Correct arcs	46.6	44.6	34.4	79.9	47.9	56.1	89.0	52.1	57.5
Reverse arcs	35.6	40.9	60.1	15.2	27.1	40.5	8.7	34.4	40.2
Additional arcs	17.8	14.5	5.5	4.9	25.0	3.4	2.3	13.5	2.3

the overall success of the PBIL+T and PBIL+BT, some cases showed relatively lower improvement. Thus, future work should aim to reveal the reasons why these discrepancies arise, even though it would be complicated by the fact that such analysis is a data-dependent task.

Acknowledgments. This research was supported by Basic Science Research through NRF funded by the Ministry of Education, Science and Technology (2011-0004113) and Kyungpook National University Research Fund, 2012.

References

1. Armananzas, R., et al.: A Review of Estimation of Distribution Algorithms in Bioinformatics. BioData Mining, vol. 1, no. 6 (2008)
2. Blanco, R., Inza, I., Larranaga, P.: Learning Bayesian Networks in the Space of Structure by Estimation of Distribution Algorithms. International Journal of Intelligent Systems, vol. 18, pp. 205–220 (2003).
3. Gosling, T., Jin, N., Tsang, E.: Population based Incremental Learning with Guided Mutation versus Genetic Algorithms: Iterated Prisoners Dilemma. IEEE Congress on Evolutionary Computation, pp. 958–965 (2005)
4. Handa, H.: Estimation of Distribution Algorithm with Mutation. LNCS, vol. 3448, pp. 112–121 (2005)
5. Pelikan, M., Sastry, K.: Initial-Population Bias in the Univariate Estimation of Distribution Algorithm. U. of Missouri.St. Louise, MEDAL Report(2009001) (2009)
6. Baluja, S.: Population-Based Incremental Learning: A Method for Integrating Genetic Search Based Function Optimization and Competitve Learning. CMU (1994)
7. Bonet, J.S.D., Isbell C.L., Viola, P.: MIMIC: Finding Optima by Estimating Probability Densities. Advances in Neural Information Processing Systems, MIT Press, Cambridge, MA, pp. 424–431 (1997)
8. Pelikan, M. et al.:BOA: The Bayesian Optimization Algorithm, Evolutionary Computation, vol. 8, no. 3, pp. 311–340 (2000)
9. Romero, T., Larranaga, P., Sierra, B.: Learning Bayesian Networks in the Space of Orderings with Estimation of Distribution Algorithms. International Journal of Pattern Recognition and Artificial Intelligence, vol. 18, no. 4, pp. 607–625 (2004)
10. Michalewicz, Z., Fogel, D.B.: How to Solve it: Modern Heuristics. Springer (2004)
11. Lauritzen, S.L., Speigelhalter, D.J.: Local Computations with Probabilities on Graphical Structures and Their Application to Expert Systems. J. of the Royal Stat. Soc.(B), vol. 50, no. 2, pp. 157–224 (1988)
12. Smith, J.W., et al.: Using the ADAP Learning Algorithm to Forecast the Onset of Diabetes Mellitus. Proceedings of the Symposium on Computer Applications and Medical Care. pp. 261–265 (1998)