Ann Oper Res
DOI 10.1007/s10479-016-2390-2

Experimental comparisons with respect to the usage of the promising relations in EDA-based causal discovery

Song Ko[1] · **Hyunki Lim**[1] · **Hoon Ko**[2] · **Dae-Won Kim**[1]

Abstract A Bayesian network is a promising probabilistic model to represent causal relations between nodes (random variables). One of the major research issue in a Bayesian network is how to infer causal relations from a dataset by constructing better heuristic learning algorithms. Many kinds of approaches were so far introduced, and estimation of distribution algorithms (EDAs) are one of the promising causal discovery algorithms. However, the performance of EDAs is considerably dependent on the quality of the first population because new individuals are reproduced from the previous populations. In this paper, we introduce a new initialization method for EDAs that extracts promising candidate causal relations based on causal scores. Then, we used the promising relations to construct a better first population and to reproduce better individuals until the learning algorithm is terminated. Experimental results show that EDAs infer a more number of correct causal relations when promising relations were used in EDA based structure learning. It means that the performance of EDAs can be improved by providing better local search space, and it was the promising relations in this paper.

Keywords Bayesian networks · Estimation of distribution algorithm · Causal discovery · Efficient learning · Promising relation

1 Introduction

A Bayesian network is a graphical representation method for the causal relations between nodes (random variables) using a directed edge. This is accomplished under the constraint of

✉ Dae-Won Kim
dwkim@cau.ac.kr

[1] School of Computer Science and Engineering, Chung-Ang University, Seoul 156-756, Republic of Korea

[2] Department of Computer Science and Engineering, Sungkyunkwan University, Seobu-ro 2066, Suwon, Republic of Korea

Published online: 08 December 2016

a directed acyclic graph (Daly et al. 2011; Neapolitan 2004). Because a causal structure gives useful inspiration for understating the characteristics of a dataset, a Bayesian network has been used as core of intelligent systems or tools such as software and medical diagnostics, and process control by NASA (Butz et al. 2009; Li et al. 2011).

The main research issue in a Bayesian network is to infer causal relations between nodes from a dataset. Among many kinds of approaches, estimation of distribution algorithms (EDAs), which are a branch of evolutionary algorithms, are one of the promising structure learning algorithm (Armañanzas et al. 2008; Blanco et al. 2003; Ding and Peng 2014; Handa 2005; Pelikan and Sastry 2009; Romero et al. 2004; Santana et al. 2013). EDAs infer the structure by improving the quality of the population, which is accomplished by generating new individuals from individuals of the previous population using three genetic operations (i.e., selection, reproduction, and evaluation).

Similar to the learning procedure of genetic algorithms (GAs), EDAs enhance the quality of the population by repeating genetic operators until terminating conditions are satisfied. Only there is one significant difference between them in that EDAs do not have a mutation operator unlike GAs. Therefore, it is more important to generate better quality of the first population compared to GAs because the typical EDAs cannot recover missed partial solutions through the rest genetic operators.

EDAs reproduce new individuals based on the probabilistic model, which is constructed using K individuals ($1 \leq K \leq M$, M: population size) among the population. Blanco et al. (2003) firstly adopted this approach to the structure learning in a Bayesian network. Several EDAs were designed based on the construction method of the probability model, for example, the univariate marginal distribution algorithm (UMDA), population-based incremental learning (PBIL), mutual information maximizing input clustering (MIMIC), and dependencies-tree-based EDA (DTEDA; Armañanzas et al. 2008; Blanco et al. 2003; Ding and Peng 2014; Pelikan and Sastry 2009; Santana et al. 2013).

In this paper, our concern is to enhance the performance of EDAs by using the promising causal relations instead of the usage of all possible relations between nodes. The search space is too extensive to explore all of possible relations during the structure learning. Our solution for that is to extract the promising relations by evaluating causalities between nodes. If two nodes have causal relations, the conditional probability distribution (CPD) of a child is significantly dependent on its parent. We called this dependency by the causality in this paper. Generally, nodes have causal relations with only a few other nodes, and they have greater causalities than others. We evaluate causality between all nodes, and then extract the promising relations based on causality scores. We use the promising relations in constructing the initial population and in reproducing new individuals during the evolution. On the contrary, typical EDAs generate an initial population under uniform probability distribution without filtering redundant relations, which are not useful in inferring correct causal relations. We conducted various experiments to show that the performance of EDAs were noticeably dependent on how the promising relations were used in EDA-based evolution. Experimental results indicate that EDAs inferred more correct causal relations when the promising relations were adopted only in the initialization and in all process of the evolution. It means that the performance of EDAs can be improved by providing better local search space, and it was the promising relations in this paper.

Algorithm 1 Framework of estimation of distribution algorithms

```
g ← 0
generate M individuals to construct initial population G(g)
while repeat until termination conditions are satisfied do
  evaluate every individual of G(g) using scoring function; i.e Bayesian Dirichlet
  select an individual subset S(g) among M individuals of G(g)
  construct probability model P(g) using S(g)
  generate new M individuals randomly based on P(g)
  select M individuals among M individuals of G(g) and newly generated M individuals for G(g + 1)
  g ← g + 1
end while
```

2 Related works

2.1 Estimation of distribution algorithms

In this section, we describe the working procedure of the classical EDAs because the proposed method is a new method to improve the performance of EDAs. The basic workflow of EDAs is identical to that of genetic algorithms (GAs) as shown in Algorithm 1, only EDAs do not take any mutation operator and use the probability model to generate new individuals instead of the crossover operator of GAs

As similar with GAs, firstly EDAs construct the initial population by randomly generating M individuals. Then, EDAs improve the quality of the population through the repetition of generic operators such as evaluation, construction of probability model, generation of new M individuals and selection of M individuals for the next generation. The individuals are evaluated using the typical scoring functions such as BD (Bayesian Dirichlet), MDL (minimum description length) and so on; in this paper we adopted BD. Then, m individuals (m is a random value where $2 \leq m \leq M$) are selected to construct the probability model; Fig. 2 is one example of the probability model. One new individual is generated under the probability model, and this procedure is repeated M times. We obtain $2 \times M$ individuals (M individuals of the previous population and newly generated M individual). Then, M promising individuals are selected for the next generation which are determined with the scoring function. These all procedures are repeated until termination conditions are met such as the number of generation, and a specific score from the scoring function.

As previously mentioned, EDAs do not take mutation operator for the simplicity of the algorithms. Therefore, some of correct relations, which are not included in the first population, cannot be restored until EDAs are terminated. Furthermore, if some of correct relations were missed in some generation, they cannot be also recovered until EDAs are terminated.

Until so far, many kinds of EDAs are introduced and they can be categorized into three groups depending how to construct the probability model; univariate model, bivariate based model, and multivariate model (Baluja 1994; Larranaga and Lozano 2002; Mühlenbein and Paass 1996; Pelikan et al. 2002). Equation 1 represents how the probability model is constructed in univariate model.

$$p(X_1, X_2, \ldots, X_n) = p(X_1)p(X_2), \ldots, p(X_n) \tag{1}$$

Based on the assumption that all nodes (X_1, ..., X_n) are independent each other, it calculates the marginal probability distribution of nodes using selected M individuals, S(g). Then, new individuals are generated based on the probability model (Baluja 1994; Mühlenbein and Paass 1996; Harik et al. 1999).

Equation 2 shows the calculating method under an assumption that there are interrelations between nodes. If the number of parent nodes of X_n, ps_{X_n}, is at most one, it is categorized into bivariate model. The probability distribution of nodes are evaluated under the condition of its one parent node, then following process is similar to univariate model (Baluja and Davies 1997; De Bonet et al. 1997; Pelikan and Mühlenbein 1999). For reference, at least one node has no parent node, for example a root node in case of the tree structure.

$$p(X_1, X_2, \ldots, X_n) = p(X_1|ps_{X_1})p(X_2|ps_{X_2}), \ldots, p(X_n) \tag{2}$$

If some of nodes have one or more parent nodes, ps_{X_n}, it is categorized into multivariate model (Etxeberria and Larranaga 1999; Pelikan et al. 2000; Pelikan 2005). Our concern in this paper is to compare the performance of EDAs between the typical approach and our approach. We select one or two methods per categorization for the performance comparisons, and evaluate the performances of EDAs based structure learning under the typical methods and under the usage of the promising relations.

2.2 Two difficulties of the structure learning

Inferring the causal structure is hard task due to two reasons; the search space and the likelihood equivalence. First of all, the number of the possible solutions can be obtained through Eq. 3, and it shows that the number of the possible structures are super–exponentially increased according to n, the number of nodes. For example, the numbers of the possible structures are 4.175e18 in $n = 10$, and the numbers are 2.344e72 in $n = 20$. Therefore, almost algorithms were designed to infer a better local optima by constructing a better heuristic algorithm.

$$f(n) = \sum_{i=1}^{n} (-1)^{i+1} \binom{n}{i} 2^{i(n-i)} f(n-i),\ f(0) = 1,\ f(1) = 1. \tag{3}$$

Another difficulty is the likelihood equivalence. If the conditional independence sets of structures are equivalent, they are called as the likelihood equivalent. Figure 1a, c are one example of the likelihood equivalence because $Node2$ and $Node3$ are independent given $Node1$. As shown in these causal structure, their causal relations are different each other. However, these structure are given equivalent scores under the typical scoring functions such as BD (Bayesian Dirichlet) scoring function (notes that $Node2$ and $Node3$ are not independent given $Node1$ in Fig. 1d; Chickering 2002; Daly et al. 2011; Yang and Chang 2002). Several scoring functions such as MIT give different scores depending on the number of states of nodes. However, it cannot be seen that the number of the state of nodes means the causal relations. Many kinds of researches in a Bayesian networks were conducted with focussing on the learning algorithm, and relatively fewer methods were introduced for the

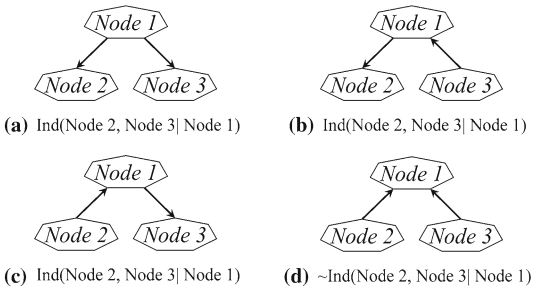


Fig. 1 **a**–**c** have same conditional independent set, Ind($Node2$, $Node3|Node1$). Otherwise, we do not say that $Node2$ and $Node3$ are conditional independent given $Node1$ in **d** structure

likelihood equivalence. In this paper, we adopted the scoring function of Ko and Kim (2014) to solve the likelihood equivalence.

3 Proposed method

It is an important issue to generate better individuals during the evolution for a better performance of EDAs. In this case, the promising relations can be a solution because EDAs reproduce better individuals by combining the promising relations with less number of redundant relations. It raises the probability of reproduction better individuals. We thought that the promising relations can be used in the initialization and in the evolution, and we designed a novel EDA based learning method based on that. The proposed method consists of three steps. Causalities between all pairs of nodes are evaluated in the first step, and then the proposed method extracts promising relations based on those causality scores. In the last step, the causal structure is inferred through EDA based structure learning using those promising relations.

3.1 Evaluation of causality between nodes

Selecting a good scoring function is no less important than selecting a good learning algorithm to obtain a better solution. In this paper, we adopted a scoring function of Ko and Kim (2014) because one of our goals is to infer the correct causal relation by handling the issue of the likelihood equivalence. We summarize the property of this scoring function for causal discovery in this section. First of all, the typical scoring functions do not distinguish causal structure between the class of the likelihood equivalence. In the case of Fig. 1a, c, all probability distribution of those structures are as follows:

$$P(Node_1, Node_2, Node_3) = P(Node_1)P(Node_2|Node_1)P(Node_3|Node_1) \quad (4)$$

All (marginal and conditional) probability distributions are equivalent and therefore the typical scoring functions give undistinguishable scores for them. Ko and Kim (2014) constructed a scoring function which evaluates causality scores between nodes from a static dataset. If nodes have causalities, a child node has a greater dependency under the others and it is represented in the conditional probability distribution (Pearl 2014). Although any correlation metric can be adopted to evaluate the conditional probability distribution, it has a limitation in separating the causal structure between the class of the likelihood equivalence.

On the contrary, the scoring function of Ko and Kim (2014) has a nonsymmetric characteristic which is the basis of the separation between the class of the likelihood equivalence. They assumed that it is sufficient to evaluate only the CPD of a child for a causal discovery, and they called the score for the conditional probability distribution as a causality score. If two nodes have causal relations, a child node is considerably dependent on its parent node and it is represented in the CPD of the child. If a given parent node is correct, then the dependency of the child shows greater than that under an incorrect parent. Based on this, Ko and Kim (2014) constructed a novel scoring function which evaluates the dependency after determining two key points on Binomial probability density function; representative point (R) and comparative point (C) (Eq. 5).

$$Dep(X_i|X_j) = \sum_{k=1}^{r_j} \{Bi(R_k) - Bi(C_k)\}, \quad \text{where } r_j \text{ is the number of instances of } X_j \quad (5)$$

Suppose there are two factors, car–start and fuel, which have causal relation. R is a score for the normal case under the causal relation and C is a score for the abnormal case as follows:

- R_1 is a score of P(start the car|fuel); this has a high probability.
- R_2 is a score of P(not start the car|not enough fuel); this has a high probability.
- C_1 is a score of P(not start the car|fuel); this has a low probability.
- C_2 is a score of P(start the car|not enough fuel); this has a low probability.

The scoring function of Ko and Kim (2014) evaluates the dependency between a pair of two nodes which is used as a basis for the differentiation between the likelihood equivalence. Relations between two nodes, $\{Node1, Node2\}$, are evaluated from $P(X_2|X_1)$ and $P(X_1|X_2)$ instead of $P(X_1)P(X_2|X_1)$ and $P(X_1)P(X_1|X_2)$; notes that $P(X_1)P(X_2|X_1) = P(X_1, X_2) = P(X_1)P(X_1|X_2)$. Based on this approach, the proposed method evaluates causalities between all pairs of nodes, $P(X_i|X_j)$, and then, constructs an n by n causal score table. In this paper, we have limited the boundary of the usage of the scoring function of Ko and Kim to evaluate causalities between nodes. BD is used to evaluate candidate structures during EDA based structure learning.

3.2 Extraction of the promising relations

It is important to extract promising relations among all possible relations to infer a better causal structure in feasible learning time. Although the number of the possible structures are super-exponentially increased according to the number of nodes, almost nodes have causal relations with a partial subset of the rest nodes. That is, we thought that the search space problem can be eased depending on how well promising causal relations are extracted.

In this paper, we extract the promising relations for efficient learning of EDAs. We obtained the causality scores via step 1. Our assumption is that the relation between nodes, which are evaluated with greater causality scores, are the promising relations while the relations, which are evaluated with lower causality scores, are redundant relations. We pick out the promising relations through the following four steps.

- *Step 1* Construct an n by n score table related to the causality. All causality scores between all pairs of nodes are calculated by the scoring function of Ko and Kim, and summarized in the score table; Fig. 2a.
- *Step 2* Filter out redundant relations. Set zero for the relations of lower causality scores, which are below the average of causality scores per node. For example, Node 3 has four candidate parents, $Nodes$ 1, 2, 3, and 5. An average of causality score of Node 3 is (8.1+1.3+4.5+2.4)/4 = 4.1. As a result, the proposed method sets the causality scores of $Node2 \rightarrow Node3$ and $Node5 \rightarrow Node3$ as zero; Fig. 2b.
- *Step 3* Convert the causality score table into the probabilistic relation table (PRT). Divide each causality score in the score table by the total score. For example, the causality score of $Node1 \rightarrow Node3$ is 8.1/(10.5+3.8+7.2+8.1+4.5+8.4+3.4 +4.2) = 0.16; Fig. 2c.
- *Step 4* Multiply n to PRT which means that an generated individual from PRT has n causal relations, statistically. Because our assumption is that each node has a few causal relations. Therefore, we set each node having one causal relation per node, statistically; Fig. 2d (notes that some probability may be greater than one. We set them as one.).

New individuals are generated based on this probability relation table (PRT); a score 1.0 means that this relation is always generated in every individuals, and a relations with zero score in the causality score table is not generated.

parent \ child	Node 1	Node 2	Node 3	Node 4	Node 5
Node 1	0	3.8	8.1	2.4	3.4
Node 2	10.5	0	1.3	1.5	1.2
Node 3	3.7	3.3	0	2.9	4.2
Node 4	1.5	7.2	4.5	0	1.5
Node 5	3.4	2.8	2.4	8.4	0.0

(a) Result of step 1

parent \ child	Node 1	Node 2	Node 3	Node 4	Node 5
Node 1	0	3.8	8.1	0	3.4
Node 2	10.5	0	0	0	0
Node 3	0	0	0	0	4.2
Node 4	0	7.2	4.5	0	0
Node 5	0	0	0	8.4	0.0

(b) Result of step 2

parent \ child	Node 1	Node 2	Node 3	Node 4	Node 5
Node 1	0	0.08	0.16	0	0.07
Node 2	0.21	0	0	0	0
Node 3	0	0	0	0	0.08
Node 4	0	0.14	0.09	0	0
Node 5	0	0	0	0.17	0

(c) Result of step 3

parent \ child	Node 1	Node 2	Node 3	Node 4	Node 5
Node 1	0	0.38	0.81	0	0.34
Node 2	1.05	0	0	0	0
Node 3	0	0	0	0	0.42
Node 4	0	0.72	0.45	0	0
Node 5	0	0	0	0.84	0

(d) Result of step 4

Fig. 2 Example with five nodes. We considered that the magnitude of the causality score is the strength of the causal relation between two nodes. We pick out the promising relations through the comparisons of causality scores

3.3 Causality discovery with the promising relations

In this section, we describe how those promising relations were used in EDAs based structure learning. The promising relations are used in two parts, the initialization process and the evolution of the EDA-based structure learning.

- *Step 5* Generate 50 (M) individuals for the initial population using the probability relation table (PRT) obtained from Step 4. For example, after generating n by n random table (RT), it compares PRT and RT. In Fig. 1d, let us suppose RT(1,3)=0.3, then $Node$ 1 is a parent of $Node$ 3 in a new individual because RT(1,3)=0.3 $<$ PM(1,3)=0.81. $Node$ 2 is parent of $Nodes$ 1 because causality scores of this relation is greater than 1, considering $0 \leq P(\cdot) \leq 1$.
- *Step 6* Construct probability model with respect to the type of EDAs such as univariate, tree, and multivariate models. The probability model is constructed by integrating selected K individuals and PRT.
- *Step 7* Evolve the quality of the population by repeating genetic operations (i.e., selection, reproduction, and evaluation) until terminating conditions are satisfied.

First of all, we generate the M (in this experimental, $M = 50$) individuals for the first generation based on the probability relation table. Some relations of zero causality scores are not generated in all individuals, while other some relations, which are evaluated with a score of one or more, are generated in all individuals. Therefore, the promising relations can be easily generated in the first population through our PRT.

Then, EDAs increase the quality of population through the generations. In this paper, we divided the usage of the promising relations in EDA-based learning; one is the usage the promising relations only in the initialization process, other is the usage the promising relations in the initialization process and in the evolution process. Equation 6 shows how the promising relations are used in constructing the probability model.

$$P(g) = S(g) \times \alpha + PRT(g) \times (1 - \alpha) \tag{6}$$

In this paper, we conducted the experiment under $\alpha = 0.5$.

4 Experimental results

4.1 Experimental design

We conducted the performance comparisons using five datasets (Table 1). We generated 10,000 cases per dataset using two softwares which were downloaded at http://norsys.com/ and http://dslpitt.org/genie/. We used four types of EDAs; UMDA (univariate model), MIMIC (bivariate model), TREE (bivariate model), and EBNA (multivariate model). The performance is evaluated based on how many causal relations were inferred by each algorithm. We compared the performance of three learning types depending on the usage of the promising relations; the typical EDA-based learning (EDA), the EDA-based learning with the initialization method using the promising relations (EDA+I), and the EDA-based learning using the promising relation in the initialization and the evolution (EDA+P). We averaged the performance after 30 repeated learning.

4.2 Edge-score based comparisons

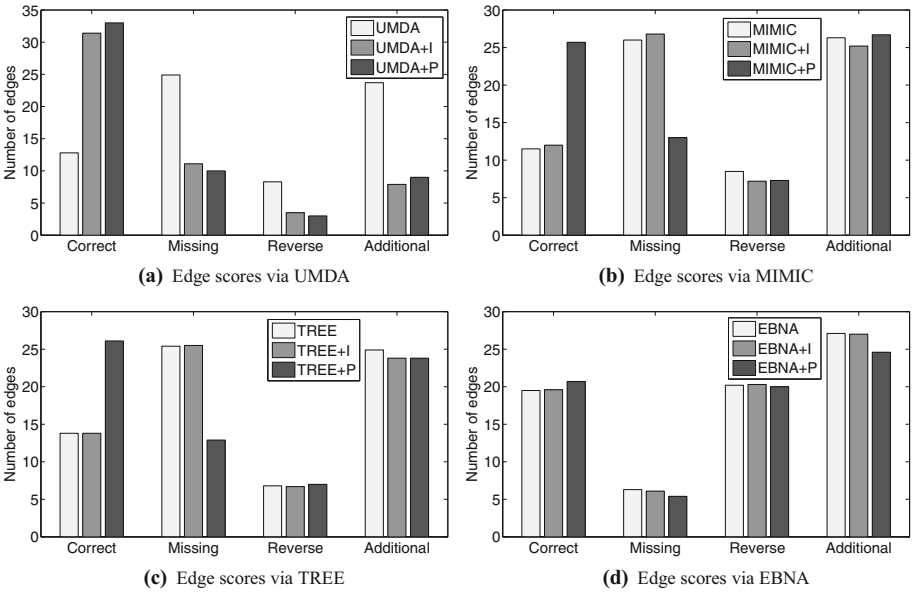
In this section, we evaluated inferred structures by counting the number of edges through the comparisons with the ground truth. We counted the number of edges after separating into four edge types. A greater number of correct edges and fewer number of missing, reverse and additional edges indicate better performance.

- **Correct edge** correctly inferred relation.
- **Missing edge** not inferred relation which is in the ground truth.
- **Reverse edge** inferred relation which indicates an opposite edge direction to the ground truth.
- **Additional edge** inferred relation that is not in the ground truth.

Figure 3 depicts the edge scores in the ALARM dataset. It showed that the typical UDDA inferred the poorest causal structure compared to the others, UMDA+I and UMDA+P. Among the entire 46 causal relations in the ALARM dataset, it inferred only 12.8 relations correctly. On the contrary, it inferred too many additional relations and missed many correct relations. The performance UMDA+I and UMDA+P inferred above 30 correct causal relations, and especially, UMDA+P inferred 33 correct causal relations. As a result, there were less missing and reverse relations than those of UMDA. In the results from MIMIC and TREE, it showed that the typical EDAs and EDAs+I have inferred causal structures of a similar performance.

Table 1 Datasets

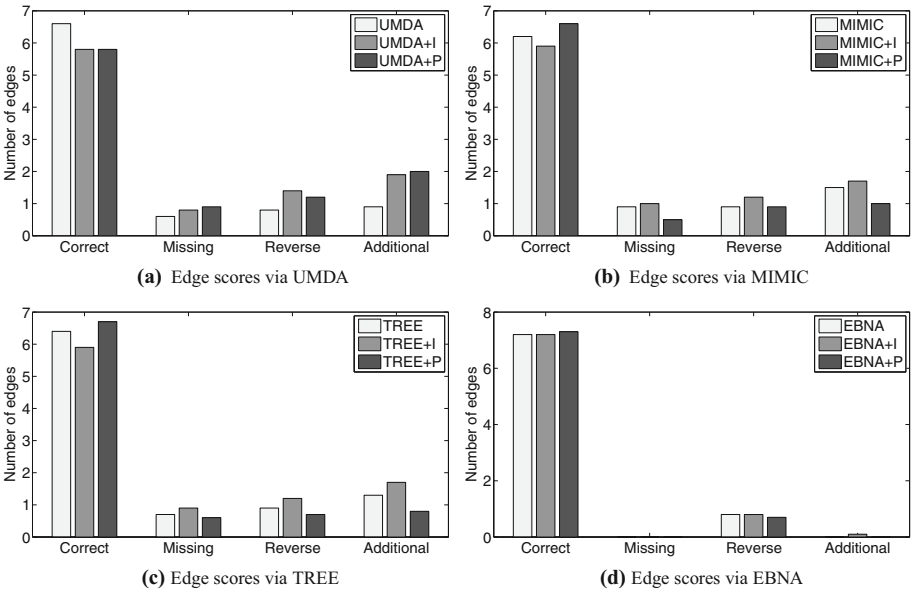
Dataset	# of nodes	# of edges	Dataset	# of nodes	# of edges
ALARM	37	46	Asia	8	8
Car	18	17	Hepar	70	120
Midway	26	38	–	–	–



(a) Edge scores via UMDA **(b)** Edge scores via MIMIC

(c) Edge scores via TREE **(d)** Edge scores via EBNA

Fig. 3 Edge scores in the ALARM dataset



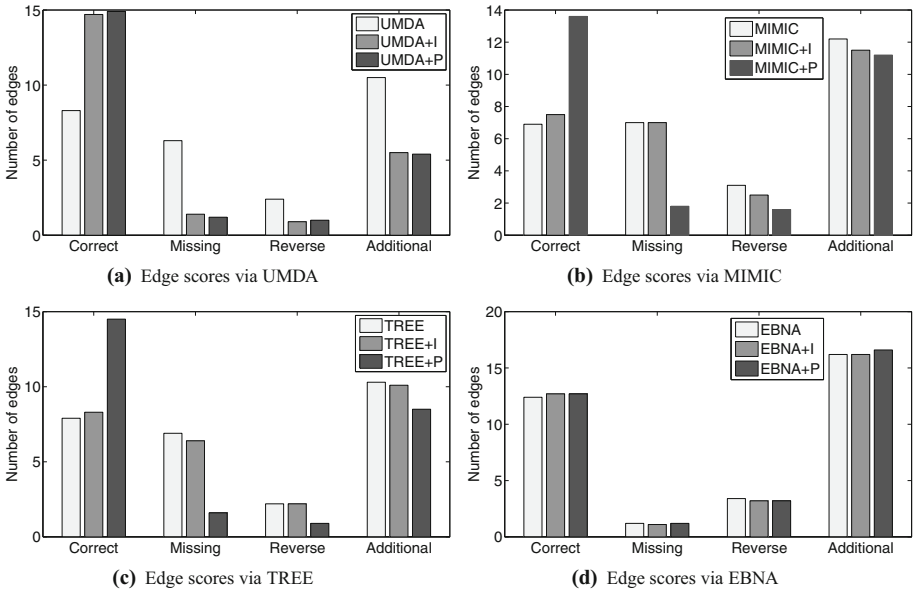
(a) Edge scores via UMDA **(b)** Edge scores via MIMIC

(c) Edge scores via TREE **(d)** Edge scores via EBNA

Fig. 4 Edge scores in the Asia dataset

Only, MIMIC+P and TREE+P have inferred above 25 correct causal relations. In the result of EBNA, all methods were showed similar performance.

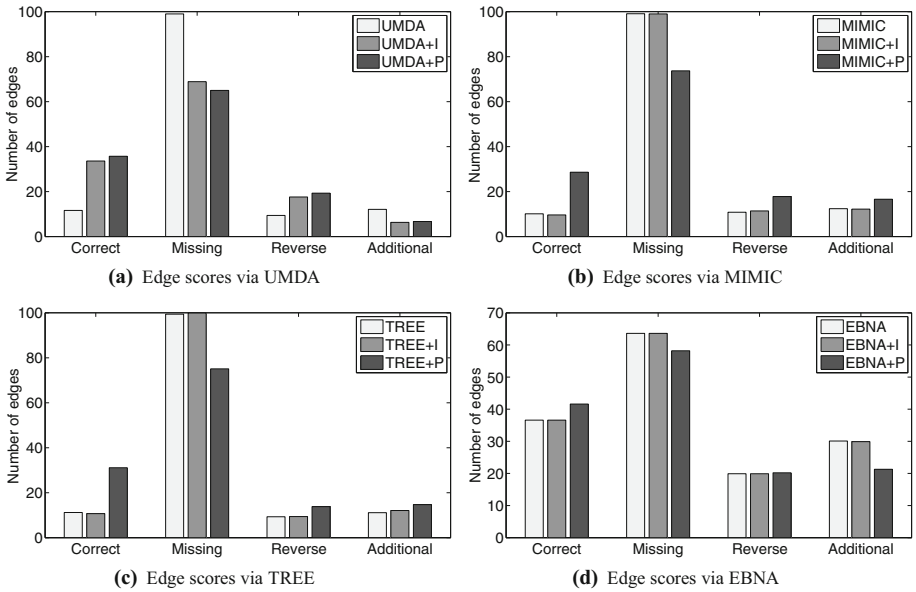
Figure 4 depicts the edge scores in the Asia dataset. It showed that all methods have inferred similar performance. Because the data size of the Asia dataset is the smallest, the



(a) Edge scores via UMDA **(b)** Edge scores via MIMIC

(c) Edge scores via TREE **(d)** Edge scores via EBNA

Fig. 5 Edge scores in the car start dataset

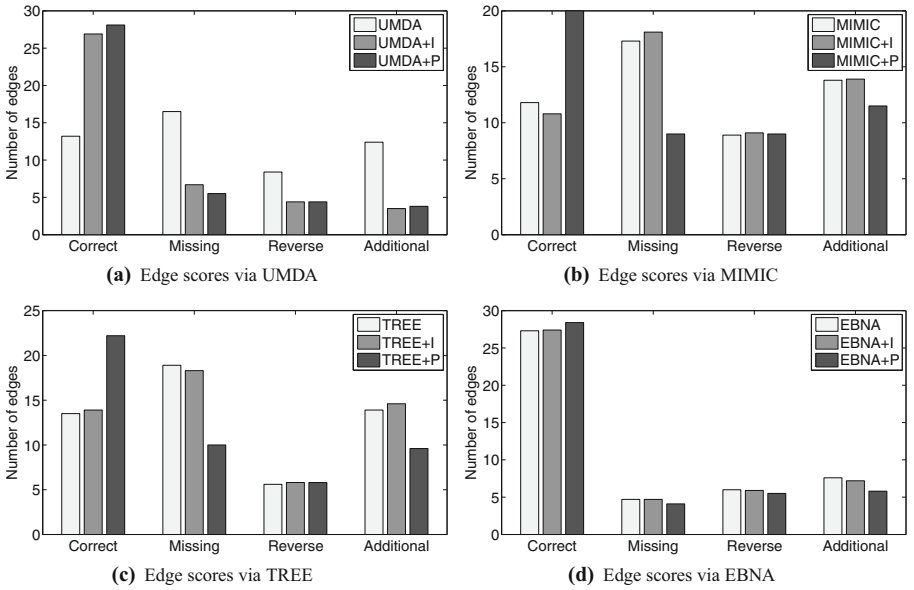


(a) Edge scores via UMDA **(b)** Edge scores via MIMIC

(c) Edge scores via TREE **(d)** Edge scores via EBNA

Fig. 6 Edge scores in the Hepar dataset

performance was not significantly dependent on the type of EDAs and the usage of the promising relations. Figure 5 depicts the edge scores in the Car dataset. We can see that the results in this dataset were similar to those of the ALARM dataset. UMDA have showed the poorest performance in the comparison of UMDAs, and UMDA+I and UMDA+P have



(a) Edge scores via UMDA

(b) Edge scores via MIMIC

(c) Edge scores via TREE

(d) Edge scores via EBNA

Fig. 7 Edge scores in the midway dataset

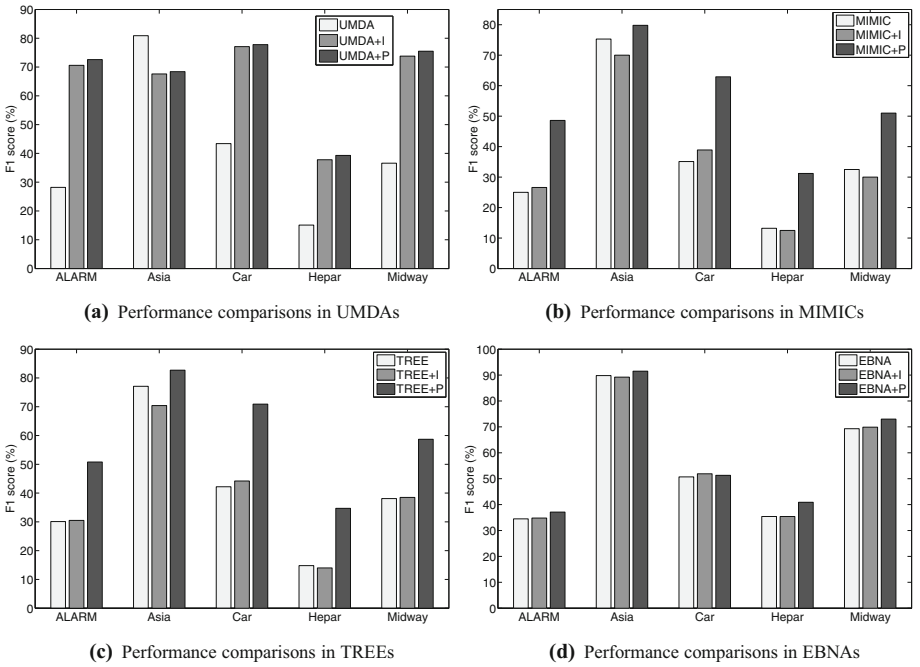
inferred the best causal structure. All EBNA based learning regardless of the usage of the promising relations showed the similar performance.

Figure 6 depicts the edge scores in the Hepar dataset. Although the promising relations were helpful to infer more number of correct causal relations, the results indicated that many causal relations were missed during the structure learning. Similar to the results in the ALARM and the Car datasets, the usage of the promising relations only in the initialization were not noticeable helpful to increase the performance to MIMIC and TREE. On the other hand, we can see that the experimental results was similarly obtained in the Midway dataset (Fig. 7). The usage of the promising relations in the initialization was sufficient to increase the performance for UMDA, while MIMIC and TREE were increased the performance when the promising relations were used in all process; the initialization process and the learning process.

4.3 F-measure based comparisons

In this section, we reevaluated edge scores with F1 scoring measure. Figure 8 showed that the performances of UMDA were increased when the promising relations were adopted except in the Asia dataset. There was little difference between two types, UDMA+I and UMDA+P, in F1 scores.

MIMIC+P and TREE+P have obtained greater scores than those of MIMIC and MIMIC+I, and than those of TREE and TREE+I, respectively. Rather, we have identified that EDAs and EDAs+I are much the same in F1 score. It indicates that many partial solutions were missed during evolutions regardless of that the promising relations were used in the initialization, and continuous providing of the promising relations were useful in the bivariate model. Finally, in the multivariate EDAs model, the usage of the promising relations was not noticeably useful. There was no significant difference in the performance between EBNA, EBNA+I and EBNA+P.



(a) Performance comparisons in UMDAs

(b) Performance comparisons in MIMICs

(c) Performance comparisons in TREEs

(d) Performance comparisons in EBNAs

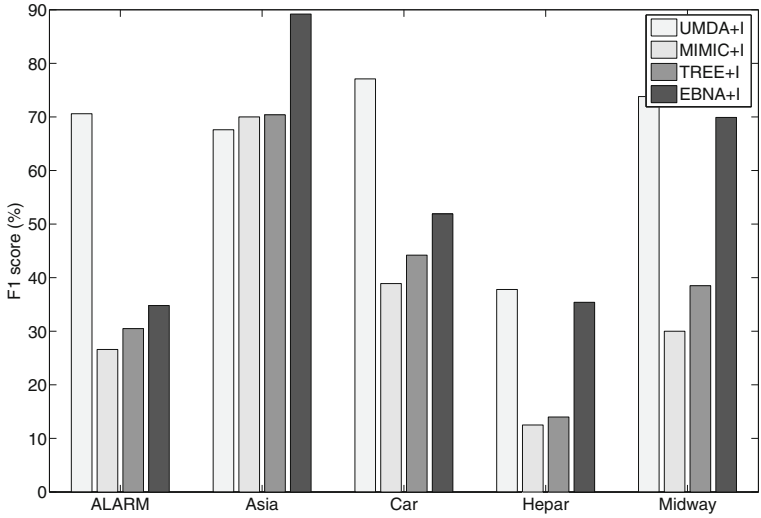
Fig. 8 F1 scores for used five dataset depending on the usage of the promising relations

4.4 EDAs-comparisons

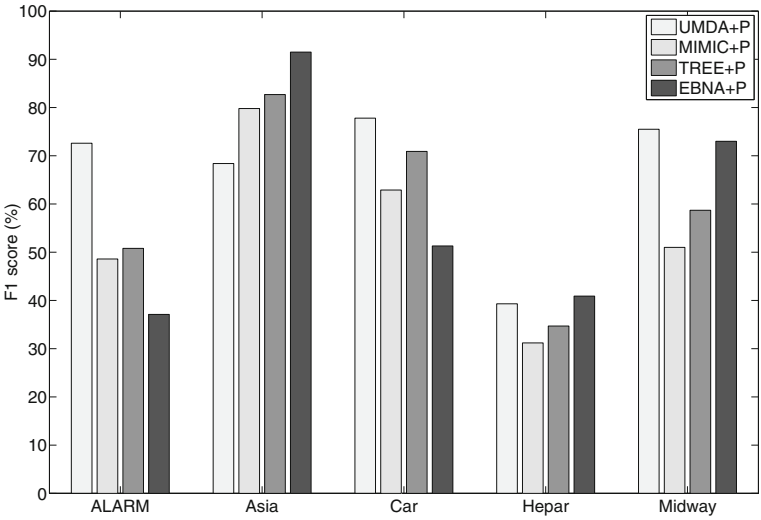
In this section, we compared the performance depending on the types of EDAs when the promising relations were used. Figure 9a depicts the results when the promising relations were used only in the initialization process and Fig. 9b depicts the results when the promising relations were used in the initialization process and in the evolution process.

First of al, when the promising relations were used only in the initialization process, generally UMDA+I and EBNA+I showed greater performances compared to the bivariate models. Among them, UMDA have showed the noticeable performance compared to the results of the others, especially in the ALARM and the Car datasets. Except for the Asia dataset, UMDA+I showed the best performance. The second best performance was obtained by EBNA+I. EBNA+I obtained the best structure in the Asia dataset. Besides, EBNA+I inferred similar performances of the structures to the results of UMDA+I, in the Hepar and the Midway datasets. On the other hand, MIMIC+I and TREE+I showed the poorest performance in this comparison.

Figure 9b depicts the performance when the promising relations were used in all procedures. It shows that UMDA+P still showed the best performance. However, it did not show significant difference in performance unlike in the results in Fig. 9a. Unlike with the results in Fig. 9a, MIMIC+P and TREE+P showed greater performances than those of EBNA+P in two datasets, the ALARM and the Car datasets. We can see that it is more useful to continue provides the promising relations during the evolution for these two models.



(a) Performance comparisons according to the types of EDAs when the promising relations were used only in the initialization process.



(b) Performance comparisons according to the types of EDAs when the promising relations were used in the initialization process and the learning process.

Fig. 9 Performance comparisons according to EDA types

5 Conclusions

In this paper, we presented the experimental comparisons of four types of EDAs with respect to the usage of the promising relations. We considered that the search space problem and the likelihood equivalent problem are eased by adopting the promising causal relations to EDAs, and the promising relations were extracted via the scoring function of Ko and Kim (2014).

We showed that the promising relations were useful to infer a better causal structure. When the promising relations were used only in the initialization method, UMDA+I and UMDA+P inferred the best performance in four datasets; ALARM, Car, Hepar, and Midway datasets. EBNA+I showed the best performance in the Asia dataset. The order of the performance when the promising relations were used in all process was as follows: UMDA+P > TREE+P > MIMIC+P > EBNA+P, and we can see that the promising relations were efficiently used in UMDA+P. Consequentially, we have showed that the performance of EDAs can be improved via better EDAs but the adjustment of the promising relations is the also efficient solution.

Although we have showed that the promising relations was increased the performance, a more practical research is needed for a more efficient EDA learning. In the large dataset, for example tens of thousands of nodes, too many nodes may be remained under the extracting method which was used in this paper. This may induce too much times to infer a causal structure. Therefore, our next research issue is to construct more efficient extracting method.

Acknowledgements This research was supported by Basic Science Research Program through the National Research Foundation of Korea (NRF) funded by the Ministry of Education (2014R1A6A3A01058174) and by Ministry of Culture, Sports and Tourism (MCST) and Korea Creative Content Agency (KOCCA) in the Culture Technology (CT) Research & Development Program 2016.

References

Armañanzas, R., Inza, I., Santana, R., Saeys, Y., Flores, J. L., Lozano, J. A., et al. (2008). A review of estimation of distribution algorithms in bioinformatics. *BioData Mining*, *1*(6), 1–12.

Baluja, S. (1994). Population-based incremental learning. A method for integrating genetic search based function optimization and competitive learning. Technical report, DTIC Document.

Baluja, S., & Davies, S. (1997). Using optimal dependency-trees for combinatorial optimization: Learning the structure of the search space. Technical report, DTIC Document.

Blanco, R., Inza, I., & Larranaga, P. (2003). Learning bayesian networks in the space of structures by estimation of distribution algorithms. *International Journal of Intelligent Systems*, *18*(2), 205–220.

Butz, C. J., Hua, S., Chen, J., & Yao, H. (2009). A simple graphical approach for understanding probabilistic inference in bayesian networks. *Information Sciences*, *179*(6), 699–716.

Chickering, D. M. (2002). Learning equivalence classes of bayesian-network structures. *The Journal of Machine Learning Research*, *2*, 445–498.

Daly, R., Shen, Q., & Aitken, S. (2011). Learning bayesian networks: Approaches and issues. *The Knowledge Engineering Review*, *26*(02), 99–157.

De Bonet, J. S., Isbell, C. L., & Viola, P. (1997). Mimic: Finding optima by estimating probability densities. In M. Mozer et al. (Eds.), *Advances in neural information processing systems* (Vol. 9, pp. 424–430).

Ding, C., & Peng, W. (2014). A robust and efficient evolutionary algorithm based on probabilistic model. *Journal of Computers*, *9*(6), 1462–1469.

Etxeberria, R., & Larranaga, P. (1999). Global optimization using bayesian networks. In *Second symposium on artificial intelligence (CIMAF-99)* (pp. 332–339). Habana, Cuba.

Handa, H. (2005). Estimation of distribution algorithms with mutation. In G. R. Raidl & J. Gottlieb (Eds.), *Evolutionary computation in combinatorial optimization* (pp. 112–121). Berlin: Springer.

Harik, G. R., Lobo, F. G., & Goldberg, D. E. (1999). The compact genetic algorithm. *IEEE Transactions on Evolutionary Computation*, *3*(4), 287–297.

Ko, S., & Kim, D.-W. (2014). An efficient node ordering method using the conditional frequency for the k2 algorithm. *Pattern Recognition Letters*, *40*, 80–87.

Larranaga, P., & Lozano, J. A. (2002). *Estimation of distribution algorithms: A new tool for evolutionary computation* (Vol. 2). Berlin: Springer.

Li, Z., Li, P., Krishnan, A., & Liu, J. (2011). Large-scale dynamic gene regulatory network inference combining differential equation models with local dynamic bayesian network analysis. *Bioinformatics*, *27*(19), 2686–2691.

Mühlenbein, H., & Paass, G. (1996). From recombination of genes to the estimation of distributions in. binary parameters. In W. Ebeling, I. Rechenberg, H.-P. Schwefel & H.-M. Voigt (Eds.), *Parallel problem solving from naturePPSN IV* (pp. 178–187). Berlin: Springer.

Neapolitan, R. E. (2004). *Learning bayesian networks*. Upper Saddle River: Pearson Prentice Hall.
Pearl, J. (2014). *Probabilistic reasoning in intelligent systems: Networks of plausible inference*. Burlington: Morgan Kaufmann.
Pelikan, M., & Mühlenbein, H. (1999). The bivariate marginal distribution algorithm. In R. Roy, T. Furuhashi & P. K. Chawdhry (Eds.), *Advances in soft computing* (pp. 521–535). London: Springer.
Pelikan, M., & Sastry, K. (2009). Initial-population bias in the univariate estimation of distribution algorithm. In *Proceedings of the 11th annual conference on genetic and evolutionary computation* (pp. 429–436). ACM.
Pelikan, M. (2005). *Hierarchical Bayesian optimization algorithm*. Berlin: Springer.
Pelikan, M., Goldberg, D. E., & Cantu-Paz, E. (2000). Linkage problem, distribution estimation, and bayesian networks. *Evolutionary Computation*, *8*(3), 311–340.
Pelikan, M., Goldberg, D. E., & Lobo, F. G. (2002). A survey of optimization by building and using probabilistic models. *Computational Optimization and Applications*, *21*(1), 5–20.
Romero, T., Larrañaga, P., & Sierra, B. (2004). Learning bayesian networks in the space of orderings with estimation of distribution algorithms. *International Journal of Pattern Recognition and Artificial Intelligence*, *18*(04), 607–625.
Santana, R., Armañanzas, R., Bielza, C., & Larrañaga, P. (2013). Network measures for information extraction in evolutionary algorithms. *International Journal of Computational Intelligence Systems*, *6*(6), 1163–1188.
Yang, S., & Chang, K.-C. (2002). Comparison of score metrics for bayesian network learning. *IEEE Transactions on Systems, Man and Cybernetics, Part A: Systems and Humans*, *32*(3), 419–428.