Journal of Advanced Scientific Research

Available online through http://www.sciensage.info/jasr

ISSN
0976-9595
Research Article

INVESTIGATION OF THE IMPACTS OF CONSTRAINT-BASED ALGORITHMS TO THE QUALITY OF BAYESIAN NETWORK STRUCTURE IN HYBRID ALGORITHMS FOR MEDICAL STUDIES

Emre Dünder*, Mehmet Ali Cengiz, Haydar Koç

Samsun Ondokuz Mayıs University Departmen of Statistics, Samsun, Turkey

**Corresponding author: emre.dunder@omu.edu.tr*

ABSTRACT

The subject of how best to present the relationships between the variables under uncertainty was crucial in recent medical studies. In this study we compared the performance of hybrid structure learning algorithms with using different constraint based methods in skeleton construction phase. We used four different medical Bayesian networks for the comparison. In conclusion, we determined the most powerful combination of learning algorithms as a two phased hybrid method for building the structure.

Keywords: Bayesian Networks, Structure Learning, Hybrid Algorithms, Medical Experiments

1. INTRODUCTION

Bayesian networks have become an important tool for decision making and statistical inference in many fields such as medicine [1], biology [2], and chemistry [3]. They provide useful representation of the joint probability distribution for a multivariate data set. Bayesian networks use graphical model to depict conditional independence among random variables in the domain and encodes the joint probability distribution [4]. A Bayesian network consists of two qualitative components: a directed acyclic graph (D.A.G) and set of conditionally probability values. In Bayesian Networks D.A.G is called the structure and conditionally probability values are called the parameters. The main task is constructing D.A.G and obtaining the parameters. These two operations are called as learning. Learning Bayesian networks is a complex task which has received a lot of attention from researchers. Structure learning algorithms have received considerable attention due to the complexity of the real world problems. Bayesian network structure can be obtained manually by expert knowledge. Without the presence of expert knowledge, structure learning algorithms provide an automated way for constructing a D.A.G given by a dataset. In recent years, there have been a lot of interests about the usage of Bayesian networks (BNs) in the medical domain. BNs are used in many areas in the medical domain such as forecasting glucose concentration [5], Detecting Asthma Exacerbations [6], analysis of an adverse drug reaction [7], diagnosis of the caring procedure for wheelchair users [8], detecting laboratory errors [9], modeling for brain longitudinal brain morphometry [10], identification of a strategic brain network [11]. Bayesian network structure learning algorithms are compared according to quality of the networks for emergency medical service [12]. The ability of BNs to model uncertainty and causal relationships among variables makes them an attractive tool in a number of medical applications. They represent causal interventions and make inferences graphically. The main advantage of using BNs to model relations in medical studies includes generating inferences with causal relationships and the conditional probabilities. Properly designed BNs can provide a valid model when any subset of the modeled variables is present. This paper explains the importance of automated structure learning algorithms in Bayesian networks for medical studies.

2. MATERIAL AND METHODS

2.1. Learning Bayesian Networks

Bayesian Networks are graphical models for reasoning under uncertainty. A Bayesian network composes of nodes and arcs. The nodes represent the random variables and the arcs represent the causal effects between the nodes. The joint distribution of the random variables are shown as a factorization of the conditional distributions

$$P(x_1, x_2, ..., x_n) = \prod_{i=1}^{n} P(x_i / Pa(x_i))$$

Where $Pa(X_i)$ is the parent of random variable x_i. The factorization of the joint distribution enables to encode the direct relations between the variables.

Learning the structure of a Bayesian network can be stated as follows: given a dataset $D = \{x_1, x_2, ..., x_n\}$ obtaining the D.A.G that best matches D. This process is achieved by using a scoring metric f which is aimed to maximize. Generally local search algorithms are implemented for learning the structure.

Hybrid approaches use both conditionally independence tests and local search algorithms. During the last years two phased algorithms are implemented such as Sparse Candidate, Max-Min Hill-Climbing.

There are mainly three approaches for structure learning: constraint-based methods, score-based methods and hybrid methods. Constraint-based methods include finding a D.A.G with using conditionally independency tests [13]. These methods try to construct a structure which reflects the relations between the variables by performing statistical hypothesis tests. Score-based methods use a scoring metric and a search method for obtaining the best structure. Generally heuristic optimization techniques are used for searching operation such as genetic algorithm [14], hill climbing [15], etc. In score-based methods the aim is to find a network which maximizes the scoring metric given a dataset.

Hybrid algorithms are based on combining both the constraint-based and score-based algorithms. Basically hybrid algorithms consist of two phases: restriction and maximization [16]. In restriction phase the related variables are determined for each variable in the network. Generally a constraint-based algorithm is implemented for the first step. In maximization phase, a score-based algorithm is implemented upon the constructed structure. The computational cost of the maximize phase is reduced by the constraint-based algorithm.

2.2. Structure Learning Methods

The quality of the network can differ with the implementation of different constraint-based and score-based algorithms in two phases during the learning process. We have selected five different constraint-based methods for the restriction phase and two different scored-based methods for the maximization phase. The description of the algorithms is presented here

- Grow Shrink (GS) algorithm utilizes the Markov blankets of the nodes by using Grow-Shrink Markov blanket algorithm [17]. This approach constructs Bayesian networks by first identifying each node Markov blankets, then connecting nodes in a maximally consistent way. The Markov blanket of a node composes of the parents, children and the common parents of the children.
- Fast Incremental Association Markov Blanket (Fast-IAMB) algorithm is a two phased constraint-based algorithm which employs a heuristic to quickly recover the Markov blankets of the nodes [18]. In first phase, the Markov blanket of each node is obtained by adding variables as long as the conditionally independence tests are reliable. In second phase, irrelevant nodes are eliminated with the shrinkage process.
- Max-min parents and children (MMPC) algorithm is based on discovering Markov blankets with two phase scheme. In first phase the related nodes are collected in the candidate Markov blanket set which is called CPC (candidate parents and children) by use of a heuristic function. CPC includes the nodes which are mostly related. In second phase for each variable the unrelated ones are excluded from CPC according to d-separation rule.
- Hill-climbing (HC) algorithm is one of the most popular score-based algorithm which implements greedy search hill-climbing local search optimization over all DAG structures. This algorithm uses both greedy hill-climbing method and scoring function such as maximizing the likelihood of the network given a dataset. Hill-climbing algorithm guarantees to construct a minimal I-map of the joint distribution under certain conditions.
- Tabu search (TABU) algorithm is another scored-based algorithm that uses tabu search strategy as an heuristic method and scoring function. Tabu search algorithm starts with a random solution and recursively select a new solution from the neighborhood of the previous one that maximally increases the scoring function [19]. The main feature of tabu search is usage of tabu list for preventing to visit forbidden solutions of D.A.G space.

In this paper we use GS, FAST.IAMB and MMPC algorithms for the restriction phase and we use HC and TABU algorithms for the maximization phase. The comparisons between the algorithms are performed with using four well-known networks. The BDeu score was used is for searching process.

3. RESULTS AND DISCUSSION

We used different structure learning algorithms for learning BNs in order to build several models. In this study we focus on hybrid structure learning methods which performs automated learning with two selection phase. In first phase the skeleton of the networks is built up on the conditionally independency relations. In second phase, a searching procedure employs to obtain the most accurate network. We investigated the effect of the algorithm in first phase to the quality of Bayesian network. We used four well-known networks about the medical field as the following: Child [20], Asia [21], Diabetes [22] and Hepar [23].

To compare the performance of the algorithms between networks, we employed the algorithms for 50 times for each network and obtained the BDeu scores. In Table 1 and Table 2, the mean and standard deviation of the scores are presented for each network and learning algorithm. According to results, when we use MMPC algorithm in restriction phase, we obtain the best score values for most of the networks. Also the best combination of hybrid algorithm composes of HC and MMPC algorithm.

Table 1. Attributes of golden standard Bayesian Networks

Network	Nodes	Arcs	Parameters	Average Markov Blanket Size
ASIA	8	8	18	2,50
CHILD	20	25	230	3,00
DIABETES	413	602	429409	3,97
HEPAR	70	1236	1453	3,51

Table 2. Mean and standart deviation of BDeu scores for hill-climbing and other algorithms

	HC+GS	HC+FAST. IAMB	HC+MMPC
HEPAR	-33777,4908 ± 227,3119	-33704,8544 ± 220,4410	-33623,9036 ± 219,5227
ASIA	-2479,98222 ± 43,0049	-2478,72208 ± 47,5862	-2474,83054 ± 48,5156
CHILD	-14508,1458 ± 278,0005	-13990,0436 ± 330,3088	-13173,45 ± 174,3032
DIABETES	-633836,442 ± 3955,3194	-597265,962 ± 3145,1210	-552422,278 ± 2049,1065

Table 3. Mean and standart deviation of BDeu scores for tabu search and other algorithms

	TABU+GS	TABU+FAST. IAMB	TABU+MMPC
HEPAR	-33808,8772 ± 203,8954	-33739,0816 ± 200,2601	-33656,2388 ± 192,6134
ASIA	-2488,60534 ± 49,4126	-2483,11394 ± 60,6399	-2485,90774 ± 49,0692
CHILD	-14449,7664 ± 251,5065	-13947,055 ± 273,9663	-13115,5542 ± 161,6034
DIABETES	-633354,532 ± 3466,2016	-597541,664 ± 3177,3995	-552641,59 ± 2066,4695

Table 4. Wilcoxon test results for Asian network

ASIA		
Algorithms	Difference of Mean Ranks	Significance
HC+FAST.IAMB - HC+GS	-0,49	0,03
HC+MMPC - HC+GS	0,39	0,01
HC+MMPC - HC+FAST.IAMB	0,88	0,00
TABU+FAST.IAMB - TABU+GS	-0,34	0,69
TABU+MMPC - TABU+GS	0,36	0,00
TABU+MMPC - TABU+FAST.IAMB	0,7	0,01

Table 5. Wilcoxon test results for Child network

CHILD		
Algorithms	Difference of Mean Ranks	Significance
HC+FAST.IAMB - HC+GS	0,88	0,00
HC+MMPC - HC+GS	2,59	0,00
HC+MMPC - HC+FAST.IAMB	1,71	0,00
TABU+FAST.IAMB - TABU+GS	0,9	0,00
TABU+MMPC - TABU+GS	2,69	0,00
TABU+MMPC - TABU+FAST.IAMB	1,79	0,00

Table 6. Wilcoxon test results for Diabetes network

DIABETES		
Algorithms	Difference of Mean Ranks	Significance
HC+FAST.IAMB - HC+GS	1,00	0,00
HC+MMPC - HC+GS	2,00	0,00
HC+MMPC - HC+FAST.IAMB	1,00	0,00
TABU+FAST.IAMB - TABU+GS	1,00	0,00
TABU+MMPC - TABU+GS	2,00	0,00
TABU+MMPC - TABU+FAST.IAMB	1,00	0,00

Table 7. Wilcoxon test results for Hepar network

HEPAR		
Algorithms	Difference of Mean Ranks	Significance
HC+FAST.IAMB - HC+GS	1,06	0,00
HC+MMPC - HC+GS	2,55	0,00
HC+MMPC - HC+FAST.IAMB	1,49	0,00
TABU+FAST.IAMB - TABU+GS	1,34	0,00
TABU+MMPC - TABU+GS	2,76	0,00
TABU+MMPC - TABU+FAST.IAMB	1,42	0,00

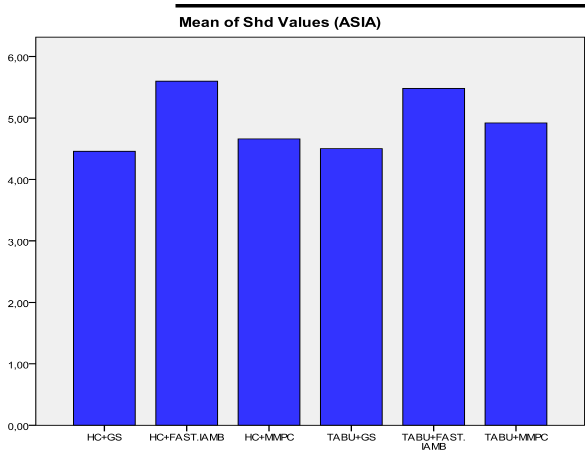

Fig. 1. Mean values of structural hamming values for Asia network.

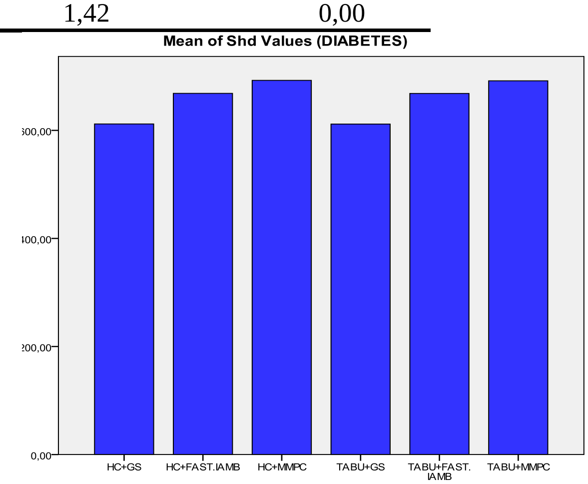

Fig. 3. Mean values of structural hamming values for Diabetes network.

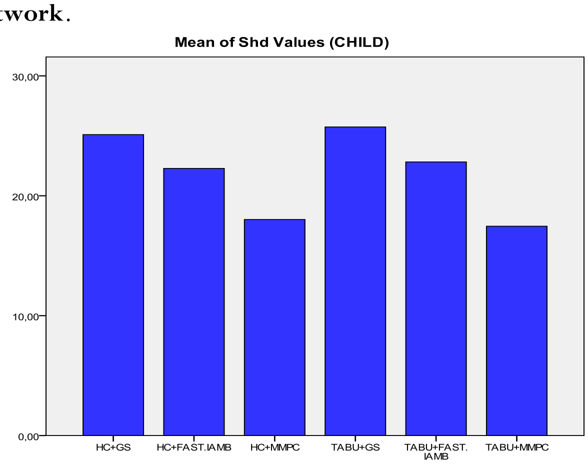

Fig. 2. Mean values of structural hamming values for Child network.

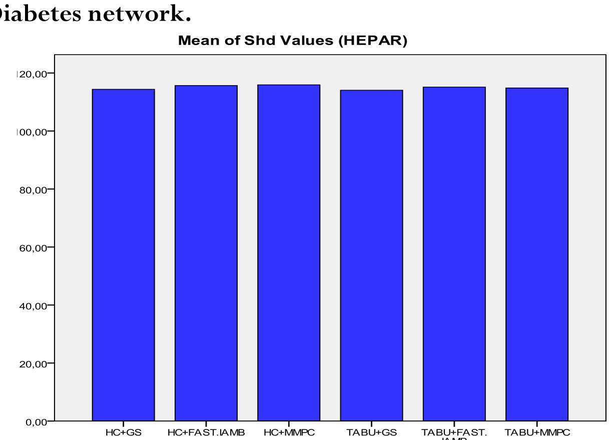

Fig. 4. Mean values of structural hamming values for Hepar network

We compared the differences between the pairs of hybrid algorithms with implementing Wilcoxon test. A summary of the analysis is shown in Table 3. Differences of the mean ranks between the pairs of hybrid algorithms and significance values are presented for the inference. According to results, for each network when we use MMPC algorithm for the restriction phase, we obtain the best qualified Bayesian network structure. Because there is a significance difference between the hybrid algorithms in case of mean ranks. The mean ranks of the hybrid algorithms which are implemented with MMPC are higher than the algorithms scores.

In order to compare the networks structures once we have computed structural hamming distances (SHD) for each learning algorithms and Bayesian networks. According to line graphs, SHD values mostly show greater performance in Bayesian networks when MMPC and GS algorithm is used. Also, the best combination of the algorithms is the score based methods and MMPC and GS algorithm.

4. CONCLUSION

In this study we investigated the effects of the selection of constraint-based algorithms in restriction phase based on two phased hybrid algorithms. We compared four different well-known Bayesian networks about medical domain for exploring the quality of networks. Descriptive statistics were computed and Wilcoxon tests were implemented for evaluating the results. From the experiments, we concluded that when we use MMPC and GS algorithm in the restriction phase, we obtain the most qualified Bayesian network structure for most of the networks. In medical studies when the Bayesian network structure is built, it is suggested to use MMPC algorithm for the first phase in hybrid learning algorithms. For the future work, different constraint-based algorithms can be used for improving the performance of hybrid algorithms.

5. REFERENCES

1. Doctor NJ, Strylewicz G. *Artificial Intelligence in Medicine*, 2010; **50:**75-82.
2. Rasmussen AL, Wang IM, Shuhart MC, Proll SC, He Y, Cristescu R et al. *Virology*, 2012; **430(1):**43-52.
3. Lau KT, Guo W, Kiernan B, Slater C, Diamond D. *Sensors and Actuators B,* 2009; **136:**242-247.
4. Ji J, Hu R, Zhang H, Liu C. *Applied Soft Computing*, 2011; **11:**3373-3384.
5. Ramoni M, Riva A, Stefanelli M, Patel V. *Artificial Intelligence in Medicine*, 1995; **7(6):**541-559.
6. Sanders DL, Aronsky D. *AMIA Annu Symp Proc*, 2006; 684-688.
7. Cowell RG, Dawid AP, Hutchinson T, Spiegelhalter DJ. *Artificial Intelligence in Medicine*, 1991; **3(5):**257-270
8. Athanasiou M, Clark JY. *Computer methods and programs in biomedicine*, 2009; **95 S:** 44-54.
9. Lippi GG, Guidi GC. *Artificial Intelligence in Medicine*, 2011; **52(3):**193.
10. Chen R, Resnick SM, Davatzikos C, Herskovits EH. *NeuroImage*, 2012; **59:**2330-2338.
11. Duering M, Gonik M, Malik R, Zieren N, Reyes S, Jouvent E. *NeuroImage*, 2013; **66:**177-183.
12. Acid S, Campos LM, Luna JMF, Rodriguez S, Rodriguez JM, Salcedo JL. *Artificial Intelligence in Medicine*, 2004; **30:**215-232.
13. Neapolitan R. Learning Bayesian Networks, Prentice Hall Series in Artificial Intelligence, 2003.
14. Larranaga P, Sierra B, Gallego MJ, Michelena MJ, Picaza JM. Learning Bayesian Networks by Genetic Algorithms: A Case Study in the Prediction of Survival in Malignant Skin Melanoma, *In Proceedings of the 6th Conference on Artificial Intelligence in Medicine in Europe*, Springer, 1997; 261-272.
15. Gamez JA, Mateo JL, Puerta JM. *Data Mining Knowledge*, 2011; **22:**106-148.
16. Scutari M. Measures of Variability for Graphical Models. PhD thesis, Universita degli Studi di Padova, Dipartimento di Scienze Statistiche, 2011.
17. Margaritis D. Learning Bayesian Network Model Structure from Data. PhD thesis, School of Computer Science, Carnegie-Mellon University, Pittsburgh, PA. CMU-CS-2003; 03-153.
18. Yaramakala S, Margaritis D. Speculative Markov Blanket Discovery for Optimal Feature Selection, *In Proceedings of the 5th IEEE International Conference on Data Mining*, 2005; 809-812.
19. Bouckaert RR. Bayesian Belief Networks: from Construction to Inference. PhD thesis, Utrecht University, The Netherlands, 1995.
20. Spiegelhalter DJ, Coewll RG. Learning in probabilistic expert systems. *In Bayesian Statistics 4* (J. M. Bernardo, J. 0. Berger, A. P. Dawid and A. F. M. Smith, eds.) Clarendon Press, Oxford, 1992; 447-466.
21. Lauritzen S, Spiegelhalter D. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 1988; **50(2):**157-224.
22. Andreassen S, Hovorka R, Benn J, Olesen KG, Carson ER. A Model-based Approach to Insulin Adjustment, *In Proceedings of the 3rd Conference on Artificial Intelligence in Medicine*, Springer-Verlag, 1991; 239-248.
23. Onisko A. Probabilistic Causal Models in Medicine: Application to Diagnosis of Liver Disorders. Ph.D. Dissertation, Institute of Biocybernetics and Biomedical Engineering, Polish Academy of Science, Warsaw, 2003.