International Journal of Engineering and Advanced Technology (IJEAT)
ISSN: 2249 – 8958 (Online), Volume-9 Issue-5, June 2020

Learning Bayesian Networks: The Combination of Scoring Function and Dataset

Saratha Sathasivam, Poh Choo Song, Jing Jie Yeap

Bayesian network (BN), a graphical model consists nodes and directed edges, which representing random variables and relationship of the corresponding random variables, respectively. The main study of Bayesian network is structural learning and parameter learning. There are score-and-search based, constraint based and hybrid based in forming the network structure. However, there are many types of scores and algorithms available in the structural learning of Bayesian network. Hence, the objective of this study is to determine the best combination of scores and algorithms for various types of datasets. Besides, the convergence of time in forming the BN structure with datasets of different sizes has been examined. Lastly, a comparison between score-and-search based and constraint based methods is made in this study. At the end of this study, it has been observed that Tabu search has the best combination with the scoring function regardless of the size of dataset. Furthermore, it has been found that when the dataset is large, the time it takes for a BN structure to converge is shorter. Last but not least, results showed that the score-and-search based algorithm performs better as compared to constraint based algorithm.

Keywords* : *Bayesian network, belief network, convergence, score, structural learning.

I. INTRODUCTION

Bayesian network (BN) comprises a set of random variables and its directed arcs which representing the conditional dependencies between nodes. It is also a directed acyclic graph (DAG). There are two important learning of BN which are structural learning and parameter learning. In the aspect of structural learning, three ways in developing the structure; which are score-and-search based, constraint based and hybrid based. On the other hand, conditional probability is studied in parameter learning, based on the network. There are numerous number of algorithms and scores arises, the question then arises as to which algorithm is the most appropriate one to be used when handling with datasets of different sizes. It is not efficient though if researchers were to use all of the algorithms and scoring methods that are available. By finding the best combination of algorithm and score (for dataset of different sizes) it enables researchers to better understand BN and they can effectively apply it when they are dealing with different types of datasets. Hence, the time spent by the researcher in obtaining the best structure network for their data is shorten and yield an efficient research. Therefore, in this study, we wish to identify a combination of scoring functions and algorithms that works best for different types of datasets. We would like to find out the algorithm(s) that has the best matching with the scoring function(s) when handling datasets of different sizes. We group the datasets into two categories, which are small datasets (<10 variables) and big datasets (≥ 10 variables).

Revised Manuscript Received on May 15, 2020.
* Correspondence Author
Saratha Sathasivam, School of Mathematical Sciences, Universiti Sains Malaysia, 11800 USM, Penang, Malaysia. Email: saratha@usm.my
Poh Choo Song*, Faculty of Science, Universiti Tunku Abdul Rahman, Perak, Malaysia. Email: songpc@utar.edu.my
Jing Jie Yeap, Faculty of Science, Universiti Tunku Abdul Rahman, Perak, Malaysia. Email: jingjieyeap@hotmail.com

II. LITERATURE REVIEW

Bayesian network (also known as belief network) is one of the probabilistic graphical models (GMs) where nodes and arcs are the major components. The random variables in BN are basically represented by nodes whereas edges denote the probabilistic dependencies among the random variables. BN is usually referred to as a DAG, which is one of the popular classes of the GMs. In other words, DAG is defined by a set of nodes and directed edges. Typically, random variables are being shown as circles with variable names labeled within the nodes. Conditional dependencies among the random variables are indicated by arrows, which connect the two nodes together [1]. The primary factors that influence Mathematical problem solving among Matriculation students in Penang are examined by adopting score-and-search based and constraint-based methods [2]. This study consists of 1312 students in total who are all from Penang Matriculation College and have enrolled in academic session 2010/2011. Each respondent is required to complete 12 questions, which are used to quantify 12 variables. All these variables are investigated in this study and thus, it is being classified as a big dataset as it consists of more than 10 variables. In conclusion, the best fitting algorithm is HC. It has the best combination as compared to all other algorithms that are used to deal with big datasets. BN has also been used to investigate the quality of sleep and health [3]. The respondents in this study are the Internet Users in Malaysia in the 10 to 50 age range. There is a total of 1316 datasets, which comprise of 20 variables (huge dataset) in the questionnaire used in this study. The highest network score goes to mmhc in three out of the five network scores.

Considering the case of a large dataset, combining mmhc with BDe, K2, and BIC yield good results in the study [3]. BN has been adopted across a broad range of disciplines where factors of floating women's income in Jiangsu are being studied by means of BN [4].

1757 respondents have been collected from the respondent for the age range 15 to 49 years old and have been migrated to other cities in China with at least three months of residence. Eight questions which will focus on eight variables are being structured in the questionnaire. Empirical evidence show that GS performs best in log-likelihood whereas HC achieves the best score in K2, BDe, AIC, and BIC. Based on prior studies, HC shows good results with all the scores used in the study, given that the dataset is small.

Reference [5] attempts to study the causal relationship between cancer susceptibility and genetic traits using BN. This study involves 1447 single-nucleotide polymorphisms (SNPs) from the cancer genes which are described by 11 variables. Due to certain constraints on this study, the exact relationship among cancer susceptibility and genetic traits remains undetermined [5]. Reference [5] concluded that score-and-search based algorithm can generate results with better precision as it attempts to obtain an optimal network by identifying all possible network structures.

Reference [6] used BN to identify the factors that affecting students in grasp on Additional Mathematics at five secondary schools in an urban area. There are 1000 respondents and 15 questions (15 variables) from the questionnaire. As compared to other scores, the main findings of the study revealed that HC and Tabu have been identified as the algorithms with the best score based on the network obtained.

It has been found that some of the researchers do not consider all the existing algorithms when conducting their research. Instead, certain algorithms are being selected from (three) structural learning methods. The question arises as to when and how to select an appropriate algorithm in the case of handling data with different number of variables. Therefore, in order to address this issue, this study aims to identify which of the combinations of scoring functions and datasets would yield the best results.

III. METHODOLOGY

A. Score-and-search Based Method

Scored-and-search based is one of the methods that can be used to learn BN [7],[8]. The score metric is used to rate each network, which reflects how well a network fits the data. It is also used to identify the network which achieves the best possible score [9]. This method works in a way in which it assigns a score to each network with the application of a heuristic search. The network with a higher score implies that it fits the data better. Hill-Climbing (HC) algorithm is a popular and widely used technique amongst heuristic search algorithms [10]. On the other hand, BN can also be constructed with Tabu search algorithm. This algorithm is introduced where it yields results that are very close to the optimal solutions. The highest possible score that is achieved indicates that the structure obtained from the algorithm is the most representative one. On the contrary, the calculation of the score of the network with a constraint based method is based on its final structure. The scores derived from all algorithms are then assigned to the networks where the corresponding scores measure and compare the networks in terms of goodness of fit.

The aforementioned algorithm, HC is a well-known search technique in score-and-search based method. The starting solution is always to begin with an empty network, and it tries to iteratively build the BN structure by trying to add, remove or reverse any possible arc [7]. The first computed score is the highest score. The iteration stops when the score of the current best solution cannot be improved with changes in arcs. HC thus, returns a solution with the best network score.

Slight modifications are implemented in HC algorithm, which is then known as Tabu search. A local search is initiated and stops whenever a local optimum is encountered. Memory structures are used in this case, to help keep the search away from recent *moves*. The modified version of the HC algorithm is then able to avoid the local optima by choosing a network that results in a minimum decrease in network score.

B. Score

Akaike Information Criterion (AIC), Bayesian Information Criterion (BIC), Bayesian Dirichlet (BDe), K2, and log-likelihood (loglik) are the few available scores in scoring a model to fit the above-mentioned algorithms. Considering the discrete case, the score functions as explained below are implemented.

The underlying cause of over-fitting by the use of log-likelihood is that the values are derived based on current data. In order to address the over-fitting issue, the complexity of the network is reduced by adding a penalty term to the log-likelihood function [11]. The penalty is written as follows:

$$Score\,(G, D) = \log\, \hat{P}\,(D \mid G) - \Delta(D, G) \qquad (1)$$

where G is the structure resulted from dataset D.

The penalty terms for BIC and AIC are shown as follows:

$$\Delta_i^{BIC} = \frac{q_i\,(r_i - 1)}{2} \ln N, \qquad \Delta_i^{AIC} = q_i\,(r_i - 1), \qquad (2)$$

where

r_i is the number of states of variable X_i,

q_i is the number of possible configuration of the parents set of variable X_i, and

N is the number of instances where variable i is in state k and j is the parent of variable i

The scores for BIC and AIC are defined as a log-likelihood function:

$$\text{BIC} = \log L(X_1, \ldots, X_v) - \Delta_i^{BIC}, \qquad (3)$$

$$\text{AIC} = \log L(X_1, \ldots, X_v) - \Delta_i^{AIC}. \qquad (4)$$

According to [12], the latter is similar to the Minimum Description Length, which has been used as a scoring function in the learning of BN [13].

The logarithm of Bayesian Dirichlet equivalent score (BDe) is a score equivalent to Dirichlet posterior density [14]. The logarithm of the K2 score (k2) is referred to as

Published By:
Blue Eyes Intelligence Engineering
& Sciences Publication

$$K2 = \prod_{i=1}^{v} K2(X_i), \quad (5)$$

$$K2(X_i) = \prod_{j=1}^{L_i} \frac{(r_i - 1)!}{(\sum_{k=1}^{R_i} n_{ijk} + r_i - 1)!} \prod_{k=1}^{R_i} n_{ijk}!. \quad (6)$$

Both scores have used the same name for the structure learning algorithm (K2 algorithm), where r_i refers to the random variables and n_{ijk} is equivalent to the above-mentioned N

C. Constraint Based Method

Learning of BNs can also be done by means of constraint based method through capturing the independence relationship, which obeys the Markov property. Constraint based algorithms are the optimized derivatives of the Inductive Causation algorithm [15]. The structure of BN is learned based on the use of conditional independence tests to find the Markov blankets of the variables. Constraint based method is found to be more effective when it comes to dealing with a large number of instances. A few of the learning algorithms which are used in the constraint based method are GS (Grow-Shrink), IAMB (Incremental Association Markov Blanket), Fast-IAMB and Inter-IAMB (Interleaved-IAMB).

Two phases are involved in the GS algorithm, which are known as the grow phase and the shrink phase. The starting solution of the grow phase is always an empty set *S*. In the case that the variables are found to be dependent on variable *X*, these variables are then added to *S* at each iteration in the grow phase. As for the shrink phase, instances will be removed from *S* if they are not in the underlying Markov network. Directions for the undirected edges are then determined according to the d-separation criterion [7].

IAMB and GS both exhibit a rather similar assumption. Nevertheless, IAMB is enhanced with a significant difference in its growing phase as compared to GS, which is the set of variables is reordered during each iteration when there is a new attribute entering the blanket. The reordering operation can be implemented with mutual information, which improves the performance of IAMB as compared to GS [16].

On top of that, Fast-IAMB is an algorithm developed to strive an equilibrium between GS and IAMB. The key idea that inspires the development of Fast-IAMB is to result in an algorithm that is able to rapidly converge to the Markov blanket as GS does. Meanwhile, Fast-IAMB should be as reliable as Inter-IAMB. Fast-IAMB is able to discover the exact Markov blanket. Moreover, its reliability depends upon on its ability to identify and eliminate the false positives in the growing phase. An approach to a faster algorithm is to use a speculative stepwise forward selection which limits the use of conditional independence tests [16].

Inter-IAMB is another variant of IAMB. The shrinking phase of Inter-IAMB reflects the underlying difference between the two algorithms, which is interleaving the grow phase with the shrink phase. Inter-IAMB relies on the forward stepwise selection with it being able to eliminate false positives in the current Markov blanket. Despite the fact that Inter-IAMB is reliable, its speed is always slower as compared to GS and IAMB [16].

D. Hybrid Based

Hybrid structural learning algorithm aggregates both score-and-search based and constraint based methods, combining the features of both independence-based and score-based methods to counter their limitations [17]. Max-Min Hill-Climbing algorithm (mmhc) is derived based on the ideas from both Max-Min Parents and Children (mmpc) algorithm as well as HC algorithm. A restrict search iteration strategy is implemented in mmpc algorithm, which aims to learn the optimal network structure in the restricted search space [18]. This algorithm accepts any combination of constraint based and score-and-search based algorithms. For instance, HC with GS, HC with IAMB, HC with Inter-IAMB and so on.

E. Data

The sample data used in this study are retrieved from two main sources, which are UCI Learning Machine Respiratory and bnlearn package which is available in RStudio software. Data are being categories as the big dataset and small dataset. Data on Nursery, Asia and Car Evaluation are being regarded as small dataset whereas Chess, Insurance and Alarm are being classified as big dataset. As for the data source, data on Nursery, Car Evaluation and Chess are obtained from UCI Machine Learning Respiratory website. The remaining three datasets are extracted from the bnlearn package. All data considered in this study are discrete. The analysis does not take into any consideration of missing values.

IV. RESULTS

A. Findings

In an effort to determine the best combination of algorithms and scores, this study examines six different data, three of each of small dataset category and big dataset category. Table I to III depict the results obtained from small datasets whilst Table IV to VI revealed the results achieved based on big dataset.

Empirical results are generated with the aids of bnlearn package in RStudio. Pairing with all the scoring functions to work on Asia (8 variables) dataset (refer to Table I), HC and Tabu topped all other algorithms with the best score. However, for the data on Car Evolution (7 variables) (Table II), Tabu performs best with AIC, BIC, BDe, and K2. Also, it is worth noting that, Fast-IAMB and Inter-IAMB perform best when combined with loglik. In essence, Tabu works best with mostly all of the scores when dealing with small datasets. Therefore, Tabu is found to be the best choice to learn BN when handling small datasets. The insights gleaned from the results in Table III implies that HC and Tabu algorithms work best on the Nursery (9 variables) dataset as it achieved the highest score when combined with the scoring functions of AIC, BIC, BDe, and K2. On the other hand, when loglik is used, Fast-IAMB and Inter-IAMB stand out, given that they have attained the highest score.

As shown in Table IV, Tabu algorithm achieved the best score with AIC, BIC, BDe and K2 when it is applied to the data on Alarm with 37 variables. On top of that, HC is recognized as the perfect combination for loglik.

Rather similar to the results obtained in Table VI. A quick look at Table VI, the scores obtained from the Insurance (27 variables) data have been recorded. Tabu is found to be the best matching for almost all the scores. Regardless of the scoring functions, Tabu scores significantly better than the rest of the algorithms with the highest score being assigned to it. Therefore, it is also recommended to use Tabu when working with large datasets.

Considering the data on Chess (37 variables), Table V presents the results obtained for each algorithm when combined with the five available algorithms. It indicates that Tabu is deemed to be the best algorithm to pair with AIC, BIC, BDe, and K2 as it attains the highest network score when paired with the aforementioned algorithms. However, the algorithm that scores the highest when in combination with loglik is HC.

In short, Tabu search is applicable to most of the scores regardless of the sizes of the datasets. To better learn the network structure, it is suggested to use score-and-search based method instead of constraint based method as it provides a better network score, disregarding the size of the dataset.

Table- I: Asia

Algorithm	Score (in negative value)				
	AIC	Bde	BIC	K2	logLik
HC	**11051.90**	**11147.65**	**11107.29**	**11109.47**	**11034.90**
Tabu	**11051.90**	**11147.65**	**11107.29**	**11109.47**	**11034.90**
GS	12480.88	12563.66	12526.50	12529.70	12466.88
Fast_IAMB	12480.88	12563.66	12526.50	12529.70	12466.88
IAMB	12480.88	12563.66	12526.50	12529.70	12466.88
Inter_IAMB	12480.88	12563.66	12526.50	12529.70	12466.88
mmhc	12032.23	12120.20	12084.37	12086.05	12016.23
rsmax2	12244.60	12326.35	12286.96	12291.37	12231.60

Table- II: Car Evolution

Algorithm	Score (in negative value)				
	AIC	Bde	BIC	K2	logLik
HC	13519.83	13615.87	13699.84	13662.12	13453.83
Tabu	**13465.65**	**13575.08**	**13686.56**	**13633.79**	13384.65
GS	13656.29	13727.99	13795.39	13776.13	13605.29
Fast_IAMB	14313.60	13857.17	17889.16	14541.41	**13002.60**
IAMB	13656.29	13727.99	13795.39	13776.13	13605.29
Inter_IAMB	14313.60	13857.17	17889.16	14541.41	**13002.60**
mmhc	13519.83	13615.87	13699.84	13662.12	13453.83
rsmax2	13656.29	13727.99	13795.39	13776.13	13605.29

Table- III: Nursery

Algorithm	Score (in negative value)				
	AIC	Bde	BIC	K2	logLik
HC	**125801.4**	**126064.9**	**126283.2**	**126159.4**	125672.4
Tabu	**125801.4**	**126064.9**	**126283.2**	**126159.4**	125672.4
GS	129111.3	129270.2	129989.0	129841.7	128876.3
Fast_IAMB	174585.3	143653.5	368268.9	143664.5	**122726.3**
IAMB	129111.3	129270.2	129989.0	129841.7	128876.3
Inter_IAMB	174585.3	143653.5	368268.9	143664.5	**122726.3**
mmhc	125851.9	126094.9	126296.4	126197.9	125732.9
rsmax2	129100.3	129226.9	129350.5	129347.0	129033.3

Table- IV: Alarm

Algorithm	Score (in negative value)				
	AIC	Bde	BIC	K2	logLik
HC	218149.6	219211.2	220761.7	219549.7	**217488.6**
Tabu	**218146.8**	**219203.3**	**220727.3**	**219522.8**	217493.8
GS	346083.9	346830.6	347474.9	346822.3	345731.9
Fast_IAMB	264630.1	265395.0	266281.9	265592.0	264212.1
IAMB	262590.9	263400.7	263875.2	263527.4	262265.9
Inter_IAMB	248718.3	249534.1	250508.4	249762.2	248265.3
mmhc	221205.6	222276.7	223450.2	222436.6	220637.6
rsmax2	335968.1	336664.4	336979.8	336598.6	335712.1

Table- V: Chess

Algorithm	Score (in negative value)				
	AIC	Bde	BIC	K2	logLik
HC	34164.99	34841.40	35451.75	35061.20	**33740.99**
Tabu	**34160.99**	**34837.53**	**35432.59**	**35048.16**	33741.99
GS	42233.28	42600.33	42488.21	42468.25	42149.28
Fast_IAMB	39483.81	38786.89	44209.04	38770.42	37926.81
IAMB	38420.06	38892.32	39121.10	38829.03	38189.06
Inter_IAMB	38564.09	38438.76	42536.68	38285.65	37255.09
mmhc	34937.18	35529.81	35962.95	35653.02	34599.18
rsmax2	42001.50	42375.00	42223.04	42228.03	41928.50

Table- VI: Insurance

Algorithm	Score (in negative value)				
	AIC	Bde	BIC	K2	logLik
HC	263109.7	264391.7	266113.0	265243.8	262349.7
Tabu	**262616.2**	**263887.0**	**26550.90**	**264702.2**	**261886.2**
GS	308446.1	309235.2	310398.3	309783.4	307952.1
Fast_IAMB	292693.2	293444.6	296953.1	294366.6	291615.2
IAMB	308446.1	309235.2	310398.3	309783.4	307952.1
Inter_IAMB	290225.1	290981.3	295073.9	292101.0	288998.1
mmhc	270798.7	272102.1	27333.20	272671.2	270144.7
rsmax2	322035.7	322712.9	322865.6	322769.6	321825.7

B. Convergence

Tabu has been recognized as the algorithm that works best with most of the scores. To put this into context, the relationship between the data size and the running time of an algorithm is further investigated.

The running time of an algorithm when combined with different scores is examined and the relationship between the data size and the running time is illustrated in Fig. 1.

As evident from Fig. 1, the computational time is directly proportional to the number of data (for Nursery). It first takes 0.11 seconds to run the Tabu search algorithm with AIC, BDe and loglik. The time it takes to run Tabu with BIC and K2 is 0.1 seconds and 0.09 seconds respectively. Once the number of instances decreases to 10000, the running time for Tabu search algorithm with AIC, BIC, and BDe scores is shortened to 0.06 seconds. The size of the dataset is further reduced, to only 9000 instances remaining, the time spent on running Tabu search algorithm with loglikelihood decreases to 0.06 seconds as well. The results remain unaltered and differences can only be observed when the number of instances is further reduced to 7000.

The results shed light on the convergence time when running Tabu search algorithm with all the available scores. In particular, the time it takes to identify a BN structure converges when the number of instances in the dataset reaches approximately 9000 instances. The running time of Tabu search algorithm is 0.06 seconds for all scores when there are 7000 instances in the dataset. The run time for Tabu decreases as the number of instances is reduced to 5000, where it takes only 0.03 seconds to identify an optimal network structure. Also, it is worth mentioning that the running time reduces to 0.03 seconds for Tabu search algorithm with all the scoring functions when the data consist of 5000 instances. Subsequently, by reducing the data to only 4000 instances, it takes only 0.01 seconds to run Tabu with AIC and BIC. As for BDe and loglik, 0.02 seconds are required and the score with the shortest run time is K2, which is less than 0.01 seconds. That being said, data with 5000 instances are adequate to construct an optimal BN structure regardless of the scoring function.

Fig. 2 gives a big picture about the time taken to run Tabu search algorithm with all the scores based on the Car Evolution data. Initially, the run time for AIC, BIC, and K2 score is 0.05 seconds whereas loglik and BDe takes only 0.04 seconds to form the network structure. It can be observed that all the scores converge with only 0.01 seconds run time. For BIC score, when the number of instances increases to 15000, the run time begins to diverge. The run time then increases right until the number of instances increases to 18000. As there are 18000 instances in the dataset, BDe turns out to be the score with the shortest run time, which is 0.04 seconds whereas all other scoring functions take 0.05 seconds to find the BN structure.

The required time to reach convergence may vary due to the fact that data size and the number of variables in each dataset are different. Take Nursery dataset as an example, it has 13000 instances and 9 variables. However, there are 7 variables in the data on Car Evolution, consisting of 18000 data. This explains why the time taken to converge varies as it might be influenced by the number of instances as well as the number of variables.

The results provide a comprehensive insight on the convergence time of the scores. The running time converges as the data size reduces. In sum, it is favourable to use any of the scoring functions with a dataset of more than 5000 instances. Efficiency can thus be improved with the use of Tabu search algorithm in the case of handling data with minimum 5000 instances and less than 10 variables.

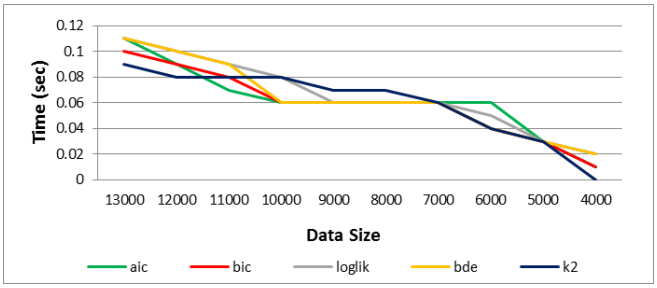

Fig. 1.Computational Time of Tabu search algorithm for Nursery data

Fig. 2.Computational Time of Tabu search algorithm for Car Evolution data

C. Comparison between score-and-search based and constraint based algorithm

Information can be garnered from the empirical results presented in Table I to VI, where score-and-search based algorithms exhibit better network score as compared to those of constraint based algorithms. Hence, it reflects the effectiveness of the both algorithms where it can be concluded that score-and-search based algorithm outperforms constraint based algorithm in identifying the optimal BN structure.

V. CONCLUSION

The main feature of BN is its ability to determine the major factors that contribute to a problem. The structure itself is also able to explain the causal relationship between the examined factors. A few of the available algorithms to perform the analysis are hill climbing, Tabu search, GS, Incremental Association Markov Blanket, Fast-IAMB, Inter-IAMB, max-min hill climbing and RSMAX2. One interesting fact about the performance of the algorithm is that Tabu algorithm is the most outstanding in terms of its appropriateness in the learning of BN. Despite the fact that other algorithms can work well in the analysis, it is believed that the best results can be delivered and the most representative structure can be produced using Tabu algorithm. This explains why researchers and practitioners are recommended to opt for Tabu algorithm when they have to handle data with different number of variables as Tabu works well with any of the scores.

The most satisfactory method to be used in finding the optimal network structure is the score-and-search based method. This study recommends further work to explore the application of hybrid based method to point out the differences between score-and-search based and constraint based method.

The difference can be evaluated based on their effectiveness, efficiency as well as accuracy. Another aspect that can be taken into account is to consider using continuous data or including missing values in future study. For the time being, numerous algorithms are being created in learning BN network structure. Due to limited availability of algorithms in the bnlearn package, the best combination of the algorithm and scoring functions is determined based on their accessibility.

ACKNOWLEDGMENT

This research is supported by the Fundamental Research Grant Scheme by Ministry of Higher Education Malaysia (203/PMATHS/6711689) and Universiti Sains Malaysia.

REFERENCES

1. I. Ben-Gal, “Bayesian networks”, Encyclopedia of Statistics in Quality & Reliability, Chichester UK: John Wiley and Sons, 2007, pp. 1-6.
2. H. C. Ong and J. S. Lim, “Identifying factors influencing mathematical problem solving among matriculation students in Penang”, Pertanika J. Soc. Sci. Hum., vol. 22(1), pp.393-408, 2014.
3. H. C. Ong, C. S. Lee and C. C. Sia, “A case study on quality of sleep and helath using Bayesian networks”, Journal of Quality Measurement and Analysis, vol. 8(2), pp. 21-26, 2012.
4. Y. Ge, C. Li and Q. Yin, “Study on factors of floating women’s income in Jiangsu province based on Bayesian networks”, Advances in Neural Network Research and Applications, vol. 67, New York: Springer, 2010, pp. 819-827.
5. C. W. Su, A. Andrew, M. Karagas, and M. E. Borsuk, “Overview of Bayesian network approaches to model gene-environment interactions and cancer susceptibility”, Bio. Data Min., vol. 6(6), pp. 1-21, 2013.
6. H. C. Ong, K. Chandrasekaran, “A Bayesian network approach to identify factors affecting learning of additional mathematics”, Journal Pendidikan Malaysia, vol. 40(2), pp. 185-192, 2015.
7. H. Cui, “Learning Bayesian network structure from data”, Master thesis, Eötvös Loránd University, 2018.
8. M. Singh, M. Valtorta, “Construction of Bayesian network structures from data: a brief survey and an efficient algorithm”, J. Approx. Reason., vol. 12, pp. 111-131, 1995.
9. Na, Y., & Yang, J., “Distributed Bayesian network structure learning”, IEEE International Symposium on Industrial Electronics (ISIE), pp. 1607-1611, 2010
10. J. Gámez, J. Mateo, and J. Puerta, “Learning Bayesian networks by hill climbing: efficient methods based on progressive restriction of the neighborhood”, Data Min. Knowl. Disc., vol. 22(1-2), pp. 106-148, 2010.
11. I. H. Witten and E. Frank, “Data Mining: Practical Machine Learning Tools and Techniques”, 2nd ed, San Francisco: Morgan Kaufmann Publishers, 2005, pp. 261-273.
12. J. Rissanen, “Modeling by shortest data description", Automatica, vol. 14(5), pp. 465-471, 2001.
13. W. Lam and F. Bacchus, “Learning Bayesian belief networks: an approach based on the MDL principle", Comput. Intell., vol. 10(3), pp. 269-293, 2000.
14. L. M. Campos, “A scoring function for learning Bayesian networks based on mutual information and conditional independence tests”, J. Mach. Learn. Res., vol. 7, 2006, pp. 2149-2187.
15. T. S. Verma and J. Pearl, “Equivalence and synthesis of causal models”, Uncertain. Artif. Intell., vol. 6, pp. 255-268, 1991.
16. S. Yaramakala, “Fast Markov blanket discovery”, Master Thesis, Iowa State University, 2004.
17. M. Scutari, “Measures of variability for graphical models”, PhD Thesis, University of Padova, 2011.
18. M. Scutari., “Learning Bayesian networks with the bnlearn R package”, J. Stat. Softw., vol. 35(3), pp. 1-22, 2010.

AUTHORS PROFILE

Saratha Sathasivam received a Bachelor Degree of Education (Science) and Master of Science (Mathematics) from Universiti Sains Malaysia and PhD from Universiti Malaya in 2001, 2003 and 2007 respectively. She was fellow under Academic Staff Training System of Universiti Sains Malaysia from 2002-2007. Currently she is serving School of Mathematical Sciences, Universiti Sains Malaysia since 2007. Her research interests mainly focus on neural network, logic programming, data mining, numerical methods and agent based modelling. She has published more than one hundreds papers on the field of neural network, logic programming etc. She is also a member of several journal’s editorial board.

Poh Choo Song received her Bachelor Degree of Computer Science (Hons) (Computational Science) and Master of Science (Statistics) from Universiti Malaysia Sarawak and Universiti Sains Malaysia in 2009 and 2011 respectively. Curretnly she is serving in Department of Physical and Mathematical Science, Faculty of Science, Universiti Tunku Abdul Rahman since 2014. Her research interests mainly focus on bayesian network, data mining and statistics. She has published several papers in the field of Bayesian network and one of her paper has won a best paper award in a conference. She is a member of The Technological Association Malayaisa (TAM) since 2018.

Jing Jie Yeap received her Bachelor Degree of Science (Hons) Statistical Computing and Operations Research from Universiti Tunku Abdul Rahman in 2016. Her final year project is in the field of Bayesian network which under Poh Choo Song’s supervision. She is dedicated in the field of data mining and statistical related topic as it is part of her research interest during her undergraduate study. Currently she is working in a manufacturing company in Penang, Malaysia for 4 years.