I.J. Wireless and Microwave Technologies, **2018, 1, 25-33**
Published Online January 2018 in MECS(http://www.mecs-press.net)
DOI: 10.5815/ijwmt.2018.01.03

***Available online at** http://www.mecs-press.net/ijwmt*

Extension of Refinement Algorithm for Manually Built Bayesian Networks Created by Domain Experts

Naveen kumar bhimagavni[a], Dr.PV Kumar[b]

[a]*Osmania University, University College of Engineering, Hyderabad and 500001, India*
[b]*Osmania University, University College of Engineering, Hyderabad and 500001, India*

Received: 10 March 2017; Accepted: 17 June 2017; Published: 08 January 2018

Abstract

Generally, Bayesian networks are constructed either from the available information or starting from a naïve Bayes. In the medical domain, some systems refine Bayesian network manually created by domain experts. However, existing techniques verify the relation of a node with every other node in the network. In our previous work, we define a Refinement algorithm that verifies the relation of a node only with the set of its independent nodes using Markov Assumption. In this work, we did propose Extension of Refinement Algorithm that uses both Markov Blanket and Markov Assumption to find the list of independent nodes and adhere to the property of considering minimal updates to the original network and proves that less number of comparisons is needed to find the best network structure.

Index Terms: Bayesian network**,** Medical Domain, Markov Assumption, Markov Blanket, Refinement Algorithm.

1. Introduction

A Bayesian network (BN) is a graph data structure that is composed of nodes and directed edges. It encodes conditional probability distributions among random variables. Nodes represent random variables and edge between them represents a statistical dependence of the child node on the parent node. Each node is associated with the conditional probability distribution of the variable given the values of its parent nodes, and this information can be used to infer the queries such as the most probable values of variables in the Bayesian Network given assignments to other variables in the Network.

* Corresponding author. Tel: +91-87 90 998 128
E-mail address: naveenkumar0206@gmail.com

Let G be a BN graph over the variables $X_1, \ldots, X_n$, each random variable X_i in the network has an associated conditional probability distribution (CPD) or local probabilistic model denoted as $P(X_i|P_aX_i)$. It captures the conditional probability of the random variable, given its parents (P_aX_i) in the graph. Probability distribution

P_B over the same space factorizes according to G if $P_B$can be represented using the chain rule for Bayesian networks as mentioned below

$$P_B(X_1, \ldots, X_n) = \prod_{i=1}^{n} P(X_i|P_aX_i)$$

The chain rule gives us a method for determining the probability of any complete assignment to the set of random variables; any entry in the joint can be computed as a product of factors, one for each variable. Each factor represents a conditional probability of the variable given its parents in the network.

1.1. Markov Assumption

The Bayesian network can also be viewed as a compact representation of a set of conditional independence assumptions about a distribution. These conditional independence assumptions are called the local Markov assumptions.

Given a BN network structure G over random variables $X_1, \ldots, X_n$, let $\text{Non-Descendants } X_i$ denote the variables in the graph that are not descendants of X_i. Then G encodes the following set of conditional independence assumptions, called the local Markov assumptions

For each variableX_i, there exists a conditional independence relation in G and it is referred as local Markov Assumption ($X_i \perp \text{Non} - \text{Descendants } X_i \ |P_aX_i$)

The local Markov assumption states that each node X_i is independent of its non-descendants given its parents.

1.2. Markov blanket

The Markov blanket for a node X_i in a Bayesian network is the set of nodes composed of X_i's parents$P_a(\ X_i)$, its children$C_h(X_i)$, and its children's other parents$P_a(Y_i)$.

Markov blanket $(X_i) \ = P_a(\ X_i) \cup C_h(X_i) \cup \ \cup_{Y \epsilon C_h(X_i)} P_a(Y_i)$

2. Related work

Generally, the network structure is learned from a dataset comprising various potentially predictive variables, some of which will be included in the network. Then the parameters of this network are trained by looking at the conditional probabilities of the variables within the dataset. Finally, an inference can be performed to predict the status of new data points that contain measurements of the same variables. As an alternate, initial Bayesian network structure can be developed by domain experts and will be refined using various techniques such as Refinement Algorithm and ExpertBayes.

Existing Refinement Algorithm considers only Markov Assumption to find the list of independent nodes. In this work, we propose Extension of Refinement Algorithm that considers the manually created Bayesian network and refines it to the best network structure with optimal time complexity while confining to the rule of making minimal updates to the initial network structure using Markov Blanket and Markov Assumption, thereby reducing the number of comparisons further.

3. New Refinement Algorithm

The updated Refinement algorithm is implemented with the set of functions developed in Octave mathematical Tool such as (i) **FindDesendents** – finds the list of dependent nodes for a given node (ii) **Updated ComputeIndependentList** – computes the list of independent nodes for a given node based on Markov Assumption and Markov Blanket (iii) **Find Cycle** – finds if there exists a cycle when an edge is added between two input nodes.

The new Refinement Algorithm extends insertion operator by prioritizing the consideration of a relation between source node S_n and destination nodeD_n, where D_n is sequentially selected from the list of independent nodes for a given source nodeS_n. First, Markov Assumption rule can be applied to a node in the network to find the list of its independent nodes. Markov Blanket rule than can be applied to a node in the network to reduce the list of its independent nodes further. Add an edge from S_n to D_n ($S_n \rightarrow D_n$) and proceed only if there will not exist any cycle; compute the present_score of the Bayesian network and compare it with the best_ score. If the present_score is greater than best score, mark present score as the best score else reverse the edge direction ($D_n \rightarrow S_n$) and repeat the above steps. This procedure can be applied to all the Source nodes in the network. best_score can be initially computed by finding accuracy with original network structure.present_score is calculated with an accuracy of the Bayesian network structure capable of inferring the outcome of a random variable.

Data:
Source Bayesian Network // initial network structure
Result:
Refined Bayesian Network

Steps:
1. Read Source Bayesian Network;
2. Validate if the input Source Bayesian Network is empty;
3. for each node (S_n) in the Source Bayesian Network
4. $L(D_n) =$ ComputeIndependentList //Computes list of Independent nodes
5. for each node (D_i) in the list $L(D_m)$
6. Add an Edge ($S_i \rightarrow D_i$)
7. If FindCycle (S_i D_i) $= 0$ then
8. Compute the present_score
9. If present_score > best_score
10. best_score = present_score

else
11. Remove an Edge $S_i \rightarrow D_i$
12. Add an Edge $D_i \rightarrow S_i$
13. repeat the steps 7, 8, 9 and 10
14. end

Algorithm 1: New Refinement Algorithm

Data:
Source Bayesian Network // initial network structure
Source Node
Result:

List of descendants of Source Node
Steps:
1. Read Source Bayesian Network;
2. Read the source node;
3. Find the list of descendants to source node;

Algorithm 2: Find Descendants

Data:
Source Bayesian Network // initial network structure
Source Node
Result:

List of Independent Nodes of Source Node S_n based on Markov Assumption

Steps:
1. Read Source Bayesian Network;
2. Read the source nodeS_n;
3. List = total nodes in BayesianNetwork;
4. Remove List = FindDescendants (S_n);
5. List = List - Remove List;

Algorithm 3: Updated ComputeIndependentList

Data:
Source Bayesian Network // initial network structure
Source Node and Destination Node
Result:
Returns Cycle = 0 or 1 based on whether cycle exists or not

Steps:
1. Read Source Bayesian Network;
2. Read the source node S_n and Destination nodeD_n;
3. List = FindDesendents (Destination nodeD_n);
4. If source node S_n is member of List then
    return 1;
  else
    return 0;

Algorithm 4: FindCycles

4. Result

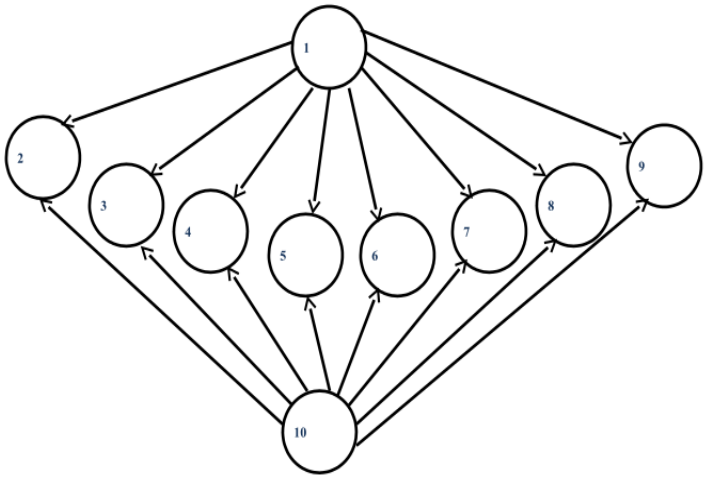

Fig.1. Initial bayesian network

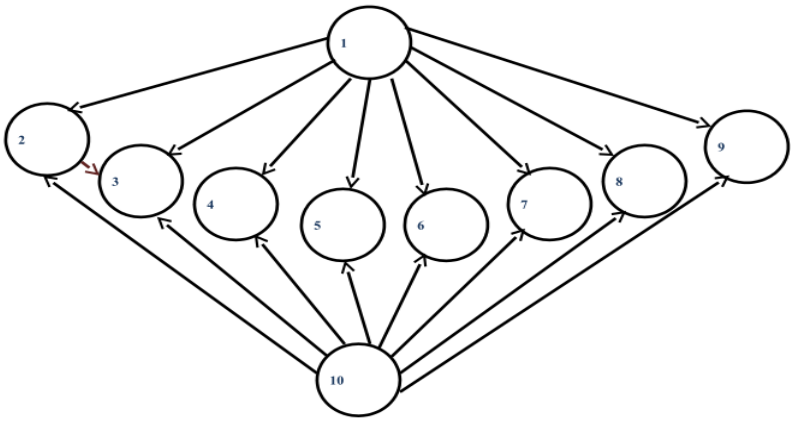

Fig.2. New refinement algorithm based Bayesian network

Table 1. Comparison table

Source node(S_i)	List of destination nodes(Dm)	New Refinement Algorithm	Refinement Algorithm
1	{10}	0	1
2	{3,4,5,6,7,8,9}	7	7
3	{4,5,6,7,8,9}	6	6
4	{5,6,7,8,9}	5	5
5	{6,7,8,9}	4	4
6	{7,8,9}	3	3
7	{8,9}	2	2
8	{9}	1	1
9	{}	0	0
10	{1}	0	1
Total number of comparisons(t)		28	30
Total number of times calculating score(2t)		56	60

When new refinement algorithm is applied to the initial Bayesian network structure as mentioned in Fig 1, *the total number of comparisons (t)* required to find the best network structure would be 28, while the existing techniques makes a total of 30 number of comparisons as mentioned in the comparison table (Table 1), another parameter *the total number of times calculating score* can be calculated by multiplying *t* with two, when considering edge direction. In addition, existing techniques calculate the score by removing and reversing edge between nodes, thereby increasing the latter parameter to 60, though the difference is small, it is considered to make a large impact on the number of nodes is increased. The resultant Bayesian network after applying new refinement algorithm is shown in Fig 2.

4.1. Result Analysis

Refinement algorithm extends existing work by pruning the number of permutations considerably to approximately less than $O\ (n^2)/2$ based on Markov Assumption and New Refinement Algorithm further reduces the number of comparisons and outputs the best score network structure. The Time complexity of new Refinement Algorithm can be estimated as mentioned below.

The worst case time complexity can be derived from three cases

Case 1: For the root nodes $(n/4)$, which don't have any parents in Bayesian network – Worst case Time complexity $T\ (C1)\ =\ O\ (n/4\ +\ 3n/4)\ =\ O\ (n)$

Case 2: For the leaf nodes $(n/4)$, which don't have any descendants in Bayesian network – Worst case Time complexity $T\ (C2)\ =\ O\ (n/4\ +\ 3n/4)\ =\ O\ (n)$

Case 3: For the non-leaf nodes (n/2), which contains both parents and descendants in Bayesian network – Worst case Time complexity $T\ (C3)\ =\ O\ (n)$

Total Time Complexity

$$T(C)\ \ =\ T\ (C1)\ +\ T\ (C2)\ +\ T\ (C3)$$

$$=\ O\ (n)\ +\ O\ (n)\ +\ O\ (n)$$

$$=\ O\ (3n)\ <\ O\ (n^2)/2$$

Recent research techniques refine the Bayesian network by considering all the possible permutations of nodes (considering the combinations of each pair of nodes in the network), which leads to the time complexity of $O\ (n^2)$ where n belongs to a number of nodes in the initial Bayesian network structure developed by domain experts.

4.2. Data and Validation

The manually built Bayesian Network is identified and collected for the widespread disease breast cancer in the medical domain; the source of data for the breast cancer was found from UCI Machine Learning Repository (Dataset Table 2). different parameters are represented as nodes in the Bayesian network such as 1.Clump Thickness, 2.Uniformity of Cell Size, 3.Uniformity of Cell Shape, 4.Marginal Adhesion, 5.Single Epithelial Cell Size, 6.Bare Nuclei, 7.Bland Chromatin, 8.Normal Nucleoli, 9.Mitoses and 10.Class: (1 for benign, 2 for malignant), the initial Bayesian network structure, manually created by domain experts was identified using the training set from the repository.

5-fold cross-validation is used to validate proposed Refinement Algorithm with consideration of dividing training set into 5 training samples being tested with test set and the results are analyzed using the Precision and Recall curve.

Table 2. Dataset

Data	Number of records	Number of Variables	Positive	Negative
Breast cancer1	400	10	140	260
Breast cancer2	299	10	71	228

5. Conclusions

When Precision and Recall analysis is performed for the 5 training samples along with test set, it is observed from the plotted graph (Fig 3) that for the same baseline Recall value, Precision values for the different training samples are in increasing order and it confirms that Refinement Algorithm reduces the probability of referring the healthy patients to the complex diagnostic tests. Table 3 describes the comparison of the proposed Refinement Algorithm with existing techniques in terms of number of comparison for the worst case scenario (naive Bayesian network structure); it can be derived from Comparison Graph (Fig 4) that number of comparisons is reduced considerably by refinement algorithm while maintaining the refinement capability intact with minor updates to the original network as mentioned in Fig 2; It is observed that time complexity is also decreased (Average reduced complexity c < 8.03%) for the various number of nodes in a given Bayesian network;

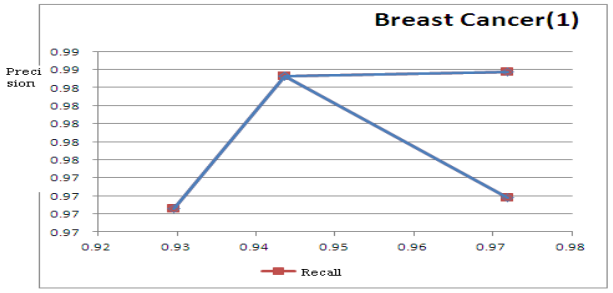

Fig.3. Precision and Recall Curves for the various thresholds

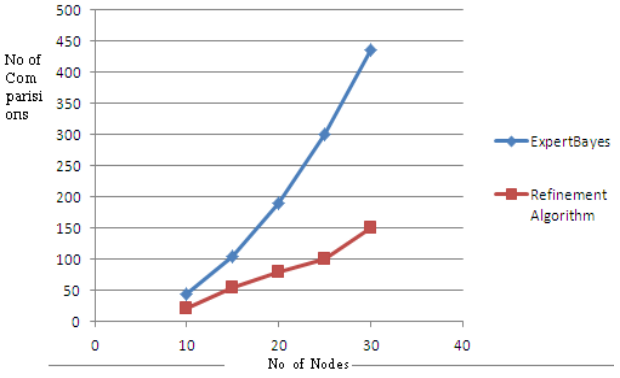

Fig.4. Comparison Graph of Refinement Algorithm with ExpertBayes

Table 3. Complexity comparison table

No of Nodes	Refinement Algorithm	New Refinement Algorithm	Reduced Complexity (%)
10	28	30	6.67
15	60	55	8.33
20	86	80	6.98
25	110	100	9.09
30	164	150	8.54

6. Future Work

We did propose and implemented new Refinement Algorithm to refine manually built a Bayesian network with reduced time complexity as compared to existing refinement techniques while confirming to the rule of making minor updates to the original network structure created by experts. Our focus is to refine this approach further to reduce the number of comparisons and another aspect is to extend this method to the Bayesian network having continuous random variables without discretizing them, thereby improving performance and estimation of classification.

Reference

[1] Ezilda Almeida, Pedro Ferreira, Tiago T. V. Vinhoza, Inez Dutra, Paulo Borges, Yirong Wu and Elizabeth Burnside. ExpertBayes: Automatically Refining Manually Built Bayesian Networks IEEE 2014 13th International Conference on Machine Learning and Applications. 2014; 362–366.

[2] Shu-bin SI1, Guan-min LIU1, Zhi-qiang CAI1, Peng XIA2. Using Bayesian Networks to Built A iagnosisand Prognosis Model for Breast Cancer; 1795-1796.

[3] Dimitris,Mar,garitis. LearningBayesianNetworkModelStructure from Data, PhD Thesis 2003;57-67.

[4] UCI Machine Learning Repository:DataSets, *archive.ics.uci.edu/ml/datasets.html?sort=nameUp&view=list.*

[5] Probabilistic Graphical Models1: Representation-Stanford Universityhttps://www.coursera.org/learn/probabilistic-graphical-models.

[6] Github.com. An implementation of Bayesian Networks Model for pure C++;2 – 6.

[7] Dr. P.J.G. Long. Introduction to Octave;4 -24.

[8] Tomasz Ku laga, Master Thesis, Jagiellonian University. The Markov Blanket Concept in Bayesian Networks and Dynamic Bayesian Networks andConvergence Assessment inGraphical Model Selection Problems. October 2006;18-20.

[9] GNU Octave, https://www.gnu.org/software/octave.

[10] Nir Friedman , Joseph Y. Halpern. A Qualitative Markov Assumption and Its Implications for Belief hange;263:1-3

[11] Henri Amuasi. Octave Programming Tutorial; 2016

[12] Daphne Koller, Nir Friedman. Probabilistic Graphical Models Principles and Techniques. The MIT Press Cambridge, Massachusetts; 2009.

Authors'Profiles

Mr. Naveen Kumar Bhimagavni, is a Research Scholar in Osmania University, did his B.Tech in CSE from JNT University, Hyderabad. Completed M.Tech in CSE from Osmania University, Hyderabad. Presently he is working in CSE department at Government Polytechnic as Lecturer.

Dr.P.V.Kumar, Professor of CSE in Osmania University Hyderabad, Completed M.Tech (CSE) and Ph.D. (CSE) degree from Osmania University, Hyderabad. He has 30 years of Teaching & R&D experience. Many students are working under him for Ph.D. He has to his credits around 56 papers in various fields of Engineering, which includes Indian, International journals, National and International conferences, He worked as Chairman BOS in OUCE and conducted various staff development programs and workshops. He is Life Member of ISTE and CSI societies.

How to cite this paper: Naveen kumar bhimagavni, PV Kumar," Extension of Refinement Algorithm for Manually Built Bayesian Networks Created by Domain Experts", International Journal of Wireless and Microwave Technologies(IJWMT), Vol.8, No.1, pp. 25-33, 2018.DOI: 10.5815/ijwmt.2018.01.03