Using Tree Augmented Naïve Bayes Classifiers to Learn Restaurant Reviews Data

Nipat Jongsawat

Faculty of Science and Technology, Rajamangala University of Technology Thanyaburi

39 Moo 1, Rangsit-Nakhonnayok Road, Thanyaburi, Pathum Thani 12110, Thailand

Email: nipat_j@rmutt.ac.th

Abstract

Wongnai is the restaurant review website. Nowadays, it becomes a top rank restaurant review website in Thailand. More importantly, lots of interesting data in Wongnai are generated when the number of reviews increase. In this paper, the data such as number of reviews, ratings, number of restaurants in reviewer's favorite list, number of photos posted, the lowest and highest price levels for food of each restaurant are manually collected from Wongnai website. A Wongnai dataset file is prepared and learnt using Tree Augmented Naïve Bayes Learning Algorithms and then a Tree Augmented Naïve Bayes graph structure is constructed. Finally, we have identified new causal relationships among variables in the constructed graph structure.

Keywords: Bayesian Networks, Tree Augmented Naïve Bayes Classifiers, Tree Augmented Naïve Bayes Learning Algorithms, Discrete Threshold

1. Introduction

Wongnai is the restaurant review website that operates with the vision of helping people discover great food around in Thailand. Launched in 2010, Wongnai has seen an amazing growth of over 400 percent this year alone. The team is bootstrapped and has just started to look for funding to further expand their business [1]. Wongnai [2] only started to get more traction in 2011 to 2012. The team has had a tough time in 2010. However, things turned to the brighter when they saw over four times growth in terms of traffic and number of users in 2012. Generally, the locals are receptive towards Wongnai. As a result, there is a very active community on Wongnai where users help each other to discover the best food or restaurant in Thailand. Wongnai now becomes a top rank restaurant review website in Thailand. More importantly, there's lots of interesting data such as number of reviews, ratings, number of restaurants in reviewer's favorite list, number of photos posted, the lowest and highest price levels for food of each restaurant in Wongnai.

When the number of reviews increase, a lot of data is generated. The ratings, number of restaurants in reviewer's favorite list, and number of photos posted will increase exponentially so that some meaningful patterns in data should be explored. Especially, the causal relationships among variables should be observed and leveraged in order to use them as a tool to aid decision making. In this paper, we are interested in determining the causal relationship between variables or factors in a directed acyclic graph or constructed model. We use the learning algorithm to create a model and then observe the causal relationships or correlations among variables. In section 2, we describe a basic concept of Tree Augmented Naïve Bayes and present some previous similar work. Tree Augmented Naïve Bayes Learning Algorithms are described in section 3. We demonstrate a learning Tree Augmented Naïve Bayes Classifier where the structure learning step is performed in section 4. In Section 5, we make a conclusion.

2. A Tree Augmented Naïve Bayes: Basic Concepts and Examples

Tree augmented naïve Bayes is a semi-naïve Bayesian Learning method. It relaxes the naïve Bayes attribute independence assumption by employing a tree structure, in which each attribute only depends on the class and one other attribute. A maximum weighted spanning tree that maximizes the likelihood of the training data is used to perform classification. Bayesian classifiers as Naïve Bayes [2] or Tree Augmented Naïve Bayes (TAN) [4] have shown excellent performance in spite of their simplicity and heavy underlying independence assumptions. In the case of TAN, a development inspired in the same idea is presented in [5], where to overcome the difficulty of exactly calculating the averaged classifier the idea of local Bayesian model averaging is introduced to calculate an approximation. In this case predictions are also improved. A tree augmented naïve Bayesian network (TAN) [3] is a Bayesian network classifier where there exists an $r \in \{1, \ldots, n\}$ such that $\Pi Xr = \{C\}$ and $\Pi Xi = \{C, Xj\}$ for all $1 \leq i \leq n$ with $i = r$. The TAN was first proposed by Friedman et al [3] to overcome the strong independence assumptions imposed by the Naïve Bayes network. In fact, the TAN is an extension of naïve Bayes which allows additional edges between the attributes of the network in order to capture correlations among them. Such correlations are however restricted to a tree structure.

In [3] an algorithm to find an optimal TAN that maximizes the log likelihood is given. The main idea is to consider a complete weighted

undirected graph, where each edge between Xi and Xj is weighted with the conditional mutual information between Xi and Xj given the class variable C. Given this, the problem reduces to determining a maximal weighted spanning tree. After computing such spanning tree, a direction has to be assigned to each edge of the tree. This is done by choosing an arbitrary attribute as the tree root and then setting the direction of all edges to be outward from it.

Jiang L. et al [6] presented a novel learning algorithm, called forest augmented naïve Bayes (FAN), by modifying the traditional TAN learning algorithm. They experimentally tested their algorithm on all the 36 data sets recommended by Weka, and compared it to naïve Bayes, SBC, TAN, and C4.4, in terms of AUC. The experimental results showed that their algorithm outperformed all the other algorithms significantly in yielding accurate rankings. Daniel L.C. Mack [7] et al described a Tree Augmented Naïve Bayes Classifier (TAN) approach to systematically extend a reference model structure using data from system operations. They compared the performance of the TAN models against a typical reference model, and demonstrate that the TAN improved classification accuracy by finding new causal links among the system monitors. They described a specific data mining approach for augmenting an existing aircraft engine reference model as an alternative to ad hoc approaches. We demonstrate the effectiveness of our work on data generated from a realistic aircraft engine simulator.

Julian et al [8] attempt to improve the performance of Web proxy cache replacement policies such as LRU and GDSF by adapting a semi naïve Bayesian learning technique. Tree Augmented Naïve Bayes classifier (TANB) classifies the web log data and predicts the classes of web objects to be revisited again future or not in the first part. In the second part, a Tree Augmented Naïve Bayes classifier is incorporated with proxy caching policies to form novel approaches known as TANB-LRU and TANBGDSF. Their proposed approach improves the performances of LRU and GDSF in terms of hit and byte hit ratio, respectively.

Lynam, T [9] describes the methods and results of applying topic modeling to 660 micronarratives collected from Australian academics/researchers, government employees , and members of the public in 2010-2011. Two tree-augmented naïve Bayes (TAN) classification models were selected to train to classify respondent Bayesian Latent Class Analysis (BLCA) Class included only discretized topic proportions (theta) for each topic and discretized sentiment score.

Alaa E. and Mahmoud F. [10] evaluate the performance of Bayesian classifier (BN) in predicting the risk of cardiovascular disease. Bayesian networks are selected as they are able to produce probability estimates rather than predictions. These estimates allow predictions to be ranked and their expected costs to be minimized. Their experimental results show that Bayesian networks with Markov blanket estimation has a superior performance on the diagnosis of cardiovascular diseases with classification accuracy of MBE model is 97.92% of test samples, while TAN and SVM models have 88.54 and 70.83%, respectively.

Jongsawat N. [11] proposed Bayesian network model based on risk taxonomy that can be used as a tool to assist in the identification of all applicable information security risks in an organization. He developed a dynamic Bayesian network model [12] to identify the Loss Event Frequency, Threat Event Frequency, and Vulnerability that reflect to information security risks of the organization. The evidence can be inferred for different time frames, where the potential attacks can be diagnosed and predicted. This model helps network analyzers understand additional information on information risk assessment of the organization based on the risk taxonomy.

3. Tree Augmented Naïve Bayes Learning Algorithms

3.1 Using Bayesian Networks for Classification

A naïve Bayes classifier learns from training data D the conditional probability of each attribute A_i given the class variable C. All variables in training data D except C are called attributes. Classification is then done by applying Bayes rule to compute the probability of C given the particular instance of $A_1...A_n$ and then predicting the class with the highest posterior probability. But in general, the strong assumption of conditional independence between the attributes given the class variable C is not realistic for real world datasets. The classification using naïve Bayes could be skewed because of the fact that it neglects the correlation between attributes in highly interrelated network. In our classifier, we start with the connections from the class variable C (Reviews) to every attribute A_i (e.g. Ranks, Ratings, Pictures, etc.). This gives the class variable a special status in the network. The connection from C to each A_i ensures that, in the learned network, the probability $P(C/ A_1...A_n)$ will take all attributes in account [5]. Thus we start with a naïve Bayes network and change it by adding edges amongst the attributes maintaining the acyclic nature of the graph. These additional edges signify a correlation

amongst variables in the structure. Thus the newly created classifier is called as augmented naïve Bayes classifier. Figure 1 shows an augmented naïve Bayesian classifier with class variable "Reviews" and three attributes Ranks, Ratings, and Pictures. Reviews is connected to all the attributes (naïve Bayes connections). There are also additional edges from Ranks to Ratings and Ranks to Pictures.

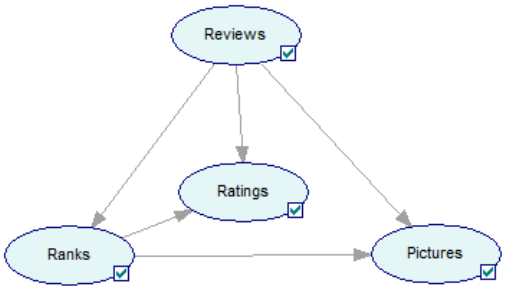

Figure 1 An augmented naïve Bayesian classifier with one class variable C and three attributes

Construction of an augmented naïve Bayesian classifier is equivalent to finding a good Bayesian network with the class variable as root. However finding the best Bayesian network from among in all the possible super exponential number of networks is a computationally intensive problem. An efficient solution of finding a useful set of edges amongst variables can be learned in polynomial time by imposing restrictions on allowable interactions amongst the variables. This results in a new network called a Tree Augmented Network or TAN.

3.2 The Construct - Tree Augmented Naïve Bayes Algorithm

The Construct-TAN algorithm [4] is shown below.

Step 1-5:

1. Compute IP (A_i ; A_j / C) between each pair of variables, where i is not equal to j

2. Build a complete undirected graph in which the vertices are the attributes A_1 ,...,A_n. Annotate the weight of an edge connecting A_i to A_j by I_P (A_i ; A_j / C)

3. Build a maximum weighted spanning tree

4. Transform the resulting undirected tree to a directed one by choosing a root variable and setting the direction of all edges to be outward from it

5. Construct a TAN model by adding a vertex labeled by C and adding an arc from C to each A_i

The Construct-TAN algorithm is a five step procedure to build a tree augmented network for a given class node C. The first step determines the weight of each edge between two nodes of the network. The weighting function is called conditional mutual information function as shown in equation (1). This function calculates the mutual information between two attributes given the value of class variable C. The CMI between the two variables X and Y can tell us whether the two variables are dependent

or not, given the value of another variable C. The dependency of the variables X and Y is determined with respect to the value of C. If CMI has a smaller value than a particular threshold t, then the two variables are conditionally independent given C. The next step constructs an undirected graph with all attributes (all nodes other than the class node) and the weights are assigned to all the edges. The next step builds a maximum-weighted spanning tree. The fourth step converts the undirected graph to directed one. The last step adds the class node to the graph.

$$I_p(X;Y/C)=\sum_{x\in X,y\in Y,c\in C} P(x,y,c)\log\frac{P(x,y/c)}{P(x/c)P(y/c)} \quad (1)$$

4. Learning the Tree Augmented Naïve Bayes Classifier from Datasets

To test our methods, we identify interesting relations between review factor and other variables, from a Wongnai rating dataset. The dataset is shown in Table 1. The data are manually collected from Wongnai website. We collected unbiased reviews of 400 popular restaurants from all over Thailand from Wongnai website. Four hundred records were analyzed in this experiment.

Table 1 Data set from Wongnai website

Ratings	Reviews	Lists	Pictures	Ranks	Min_Price	Max_Price
84.0	76.0	813.0	397.0	34.0	251.0	500.0
174.0	150.0	1400.0	2407.0	6.0	501.0	1000.0
195.0	155.0	1200.0	1050.0	7.0	101.0	250.0
417.0	378.0	2900.0	2403.0	2.0	101.0	250.0
49.0	49.0	359.0	169.0	36.0	501.0	1000.0
46.0	44.0	435.0	156.0	27.0	101.0	250.0
118.0	128.0	763.0	792.0	9.0	251.0	500.0
73.0	64.0	441.0	231.0	25.0	101.0	250.0
72.0	69.0	286.0	394.0	2.0	101.0	250.0
25.0	21.0	89.0	202.0	13.0	101.0	250.0
69.0	76.0	505.0	567.0	1.0	501.0	1000.0
173.0	121.0	1400.0	1782.0	21.0	501.0	1000.0
139.0	113.0	713.0	1028.0	39.0	251.0	500.0
343.0	282.0	3200.0	1538.0	1.0	501.0	1000.0
73.0	60.0	552.0	526.0	16.0	251.0	500.0
133.0	128.0	1000.0	1069.0	7.0	501.0	1000.0
31.0	30.0	160.0	325.0	300.0	251.0	1000.0
128.0	110.0	1700.0	759.0	4.0	501.0	1000.0
69.0	65.0	407.0	646.0	1.0	251.0	500.0
143.0	124.0	1600.0	803.0	9.0	251.0	500.0
7.0	5.0	243.0	75.0	7.0	50.0	100.0
70.0	60.0	609.0	645.0	2.0	251.0	500.0
211.0	181.0	1200.0	1026.0	8.0	101.0	250.0
24.0	23.0	267.0	140.0	299.0	101.0	250.0
54.0	49.0	397.0	366.0	45.0	251.0	500.0
138.0	126.0	1500.0	795.0	12.0	251.0	500.0
29.0	27.0	249.0	113.0	262.0	101.0	250.0
73.0	70.0	469.0	601.0	3.0	251.0	500.0
62.0	55.0	976.0	292.0	41.0	101.0	250.0
36.0	28.0	394.0	277.0	1.0	50.0	100.0
88.0	77.0	1100.0	878.0	19.0	251.0	500.0
56.0	49.0	719.0	355.0	12.0	101.0	250.0
60.0	46.0	908.0	461.0	40.0	101.0	250.0
35.0	29.0	505.0	295.0	18.0	50.0	100.0
144.0	135.0	1000.0	603.0	18.0	101.0	250.0
26.0	22.0	647.0	190.0	16.0	101.0	250.0
91.0	82.0	997.0	571.0	9.0	101.0	250.0

Each record represents one restaurant that is in the list of top restaurant from 2015-2017 (voted by the unbiased reviewers) of Wongnai website. Ratings: reviewers or gourmets give reviews and ratings to the restaurant. The total rating score of each restaurant was collected. Reviews means the number of reviews of that restaurant. Lists mean the number of times that reviewers put that restaurant on their favorite list. Pictures mean the number of photos of food and drink menus that are posted on the Wongnai website for each restaurant. Ranks mean the ranking of that restaurant (voted by reviewers). Min_Price means the lowest price level for food of each restaurant. Max_Price means the highest

price level for food of each restaurant. Both min_price and max_price were collected and suggested by the reviewers.

We examined the learning of the Tree Augmented Naïve Bayes Classifier from data sets using Smile [13], [14]. SMILE is the software library for performing Bayesian inference, written in C++, available in compiled form for a variety of platforms. Five steps were performed.

1. Compute the mutual information between each pair of attributes
2. Build a complete undirected graph in which the vertices are the attributes n variables. The edges are weighted according to the pairwise mutual information
3. Build a maximum weighted spanning tree
4. Transform the resulting undirected graph to a directed graph by selecting the class variable as the root node and setting the direction of all edges outward from it
5. Construct a TAN model by adding an arc from the class variable to all other variables

The data were learnt and the Bayesian models were constructed. A tree augmented naïve Bayes model is constructed as shown in Figure 2.

The algorithm parameters are as follows:

Learning Algorithm: Tree Augmented Naïve Bayes

Discrete Threshold: 37

Class Variable: Review

Seed: 0

Max Time (Seconds): 0

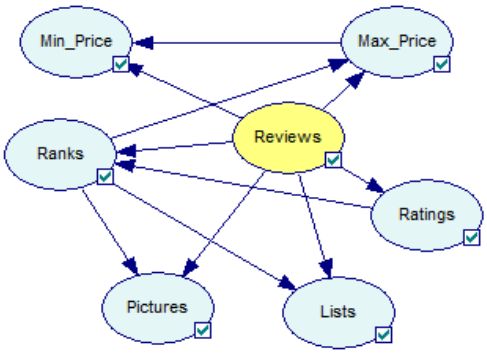

Figure 2 Tree Augmented Naïve Bayes model (Seed: 0; Max Time (Seconds): 0)

Table 2 Relationship between parent node and child nodes (Seed: 0; Max Time (Seconds): 0))

Parent Nodes	Child Nodes (is influenced)
Reviews	Min_Price Max_Price Ratings Lists Pictures Ranks
Min_Price	-
Max_Price	Min_Price
Ratings	Ranks
Pictures	-
Ranks	Max_Price Lists Pictures

Table 2 shows the relationship between parent node and child nodes (Seed: 0; Max Time (Seconds): 0).

Figure 3 shows a tree augmented naïve Bayes model when the seed parameter is equal to 1 with maximum execution time of 1 second.

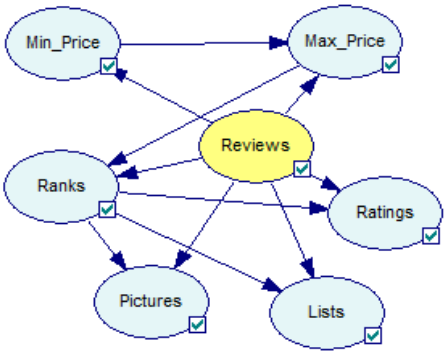

Figure 3 Tree Augmented Naïve Bayes model (Seed: 1; Max Time (Seconds): 1)

Table 3 shows the relationship between parent node and child nodes (Seed: 1; Max Time (Seconds): 1). Min_Price has a significant influence on Max_Price and Ranks influence on Ratings. We can observe that the direction of the relationship between the two pairs is in the opposite direction when compared to the causal relationship occurred in Figure 2.

Table 3 Relationship between parent node and child nodes (Seed: 1; Max Time (Seconds): 1))

Parent Nodes	Child Nodes (is influenced)
Reviews	Min_Price Max_Price Ratings Lists Pictures Ranks
Min_Price	Max_Price
Max_Price	Ranks
Ratings	-
Pictures	-
Ranks	Ratings Lists Pictures

Figure 4 shows a tree augmented naïve Bayes model when the seed parameter is equal to 2 with maximum execution time of 2 seconds.

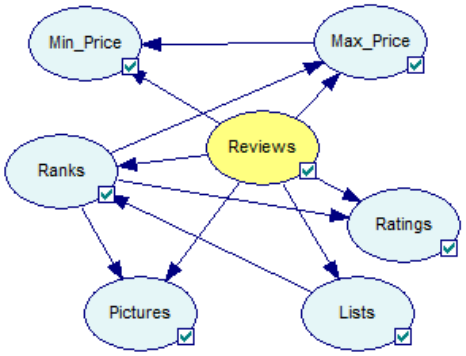

Figure 4 Tree Augmented Naïve Bayes model (Seed: 2; Max Time (Seconds): 2)

Table 3 shows the relationship between parent node and child nodes (Seed: 1; Max Time (Seconds): 1). Max_Price has a significant influence on Min_Price, Ranks influence on Max_Price, and Lists influence on Ranks. We can observe that the direction of the relationship between the three pairs is in the opposite direction when compared to the causal relationship occurred in Figure 3.

Table 4 Relationship between parent node and child nodes (Seed: 2; Max Time (Seconds): 2))

Parent Nodes	Child Nodes (is influenced)
Reviews	Min_Price Max_Price Ratings Lists Pictures Ranks
Min_Price	-
Max_Price	Min_Price
Ratings	-
Lists	Ranks
Pictures	-
Ranks	Max_Price

Figure 5 shows a tree augmented naïve Bayes model when the seed parameter is equal to 3 with maximum execution time of 3 seconds and Figure 6 shows a tree augmented naïve Bayes model when the seed parameter is equal to 4 with maximum execution time of 4 seconds. We can observe that two constructed models are the same. The node “Reviews” connects to every node. The node “Ranks” has a strong influence on the nodes “Max_Price”, “Ratings”, “List”, and “Pictures”.

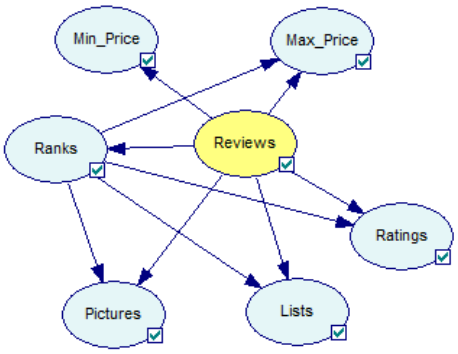

Figure 5 Tree Augmented Naïve Bayes model (Seed: 3; Max Time (Seconds): 3)

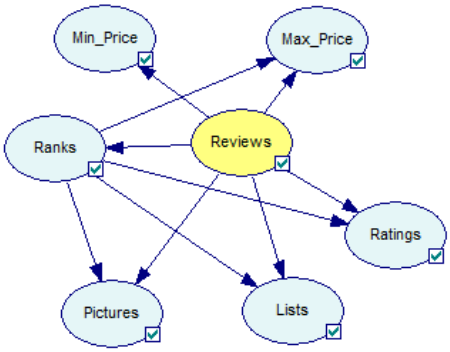

Figure 6 Tree Augmented Naïve Bayes model (Seed: 4; Max Time (Seconds): 4)

5. Conclusion

When a Wongnai dataset file is loaded and learnt using Tree Augmented Naïve Bayes Learning Algorithms, a Tree Augmented Naïve Bayes graph structure is constructed. The node "Reviews" is selected as the class variable that connects to every node for the TAN structure learning algorithm because reviews is the key factor that cause a given price level, even the lowest and highest price. Gourmets review restaurants and then assign numerical ratings to restaurants. When gourmets reviews the restaurant, they will make decisions whether to put them on their favorite list. The number of reviews may increase the number of photos shared online and the number of reviews may increase the probability of getting a higher restaurant rank. Ranks have a significant impact on the defined maximum price, ratings, lists, and pictures. Bayesian classifiers such as Naïve Bayes or Tree Augmented Naïve Bayes (TAN) have shown excellent performance given their simplicity and heavy underlying independence assumptions. In this paper, we obtain the TAN model for the Wongnai dataset. The TAN model captures the inherent dependency among the attributes in dataset. We can get some intuitive understanding about the attribute relationship in the model.

Acknowledgements

The authors would like to thank the Decision Systems Laboratory, University of Pittsburgh for supporting DSS engine (Smile), documents, and source file of the engines. All necessary files and documentations have been obtained from the Decision Systems Laboratory's web site. It is available at http://support.bayesfusion.com

References

[1] https://e27.co/finding-amazing-food-with-wongnai-thailands-yelp

[2] https://www.wongnai.com

[3] Langley, P., & Sage, S. (1994). Induction of selective bayesian classifiers. In de M`antaras, R. L., & Poole, D. (Eds.), Proceedings of the Tenth Conference on Uncertainty in Artificial Intelligence, pp. 399–406. San Francisco, CA: Morgan Kaufmann.

[4] Friedman, N., Geiger, D., & Goldszmidt, M. (1997). Bayesian network classifiers. Machine Learning, 29, 131–163.

[5] Cerquides, J. (1999a). Applying General Bayesian Techniques to Improve TAN Induction. In Proceedings of the International Conference on Knowledge Discovery and Data Mining, KDD99.

[6] Jiang L., Zhang H., Cai Z., Su J. (2005). Learning Tree Augmented Naïve Bayes for Ranking. In: Zhou L., Ooi B.C., Meng X. (eds) Database Systems for Advanced Applications. DASFAA 2005. Lecture Notes in Computer Science, vol 3453. Springer, Berlin, Heidelberg.

[7] Mack, D. L. C., G. Biswas, X. Koutsoukos, and D. Mylaraswamy (2011). Using Tree Augmented Naïve Bayes Classifiers to Improve Engine Fault Models. Uncertainty in Artificial Intelligence: Bayesian Modeling Applications Workshop. Barcelona, Spain.

[8] Julian B.P., Sagayaraj F.F., Muruganantham U. Improving the Performance of a Proxy Cache Using Tree Augmented Naïve Bayes Classifier. International Conference on Information and Communication Technologies (ICICT 2014), Procedia, Computer Science 46 (2015) 184 – 193.

[9] Lynam, T. (2016). Exploring social representations of adapting to climate change using topic modeling and Bayesian networks. Ecology and Society 21(4):16.

[10] Alaa E. and Mahmoud F. (2015). Diagnosis of Cardiovascular Diseases with Bayesian Classifiers. Journal of Computer Science. 11 (2): 274.282, DOI: 10.3844/ jcssp. 2015. 274. 282.

[11] Jongsawat N., Decharoenchitpong J., and Wuttidittachotti P. (2015). Development of a Bayesian Network Model for Information Security based on Risk Taxonomy. Engineering Journal of Siam University, 16(1), 36–46.

[12] Jongsawat N. (2016). Dynamic Bayesian Networks for Information Security. Engineering Journal of Siam University, 17(1), 40–51.

[13] https://www.bayesfusion.com

[14] http://support.bayesfusion.com/docs/SMILE