Paper 1312-2017

Construction of a Disease Network and a Prediction Model for Dementia

JinWoo Cho, SungKyunKwan University; YoungOh Seo, SungKyunKwan University;
JinYong Eom, Korea University

ABSTRACT

Regarding a human disease network, most of studies have estimated the associations of disorders primarily with gene or protein information. Those studies, however, have some difficulties in the data, because of massive volume of the data and huge computational cost. Instead, we constructed a human disease network that can describe the associations between diseases, using the claim data of Korean Health insurance. Through several statistical analyses, we show the applicability and suitability of the disease network. Furthermore, we have developed a statistical model that can predict a prevalence rate for dementia by utilizing significant associations of the network in a statistical perspective.

The source of the data is Korean HIRA (Health Insurance Review and Assessment), Korean organization that manages the overall claims of health insurance. The data is a random sample of recent 5 years (January 2010 ~ December 2014) of the whole claims, so that it contains around 1 million patients' detailed records and their sex and age. Moreover, among diseases registered in disorders exclusively. Tools that our study have used are 'SAS® Enterprise Guide® (EG)', 'SAS® Enterprise Miner™ (EM)' and 'SAS® Visual Analytics (VA)' for data processing, model construction and visualization respectively. In this study, only main diseases for each patient were extracted, so that we were able to minimize spurious associations triggered by frequencies of diseases. Models for the network are "Logistic Regression Model' and 'Bayesian Network Model' that includes 'Tree Augmented Naïve Bayes (TAN)' and 'Bayesian network Augmented Naïve Bayes (BAN).' In addition, to construct the prediction models, 'Logistic Regression', 'Decision Tree' and 'Support Vector Machine (SVM)' were used.

In conclusion, the results are quite plausible to explain the real associations between diseases. Moreover, the network has shown distinctive visualizations for each data set which is stratified by age and sex. By the stratifications, concrete and sophisticated description can be available. Especially, TAN and BAN have strong power as an explanation tool of the relationship. To illustrate, TAN has been remarkable in grasping those overall relationships of diseases through visual clustering. BAN with input and output nodes, on the other hand, has been powerful to figure out whether a particular disease is a symptom or is a cause of others. We have applied not only existing values in the raw data to the model, but also weighted variables which implied the associations we had found out in the network. As a result, in case of male patients who have had surgeries, test misclassification error rates of our chosen model is 5.57% and that of the male patients with no surgery before is 3.29%. Meanwhile, these corresponding rates are 5.45% and 9.73% for female. It can be explained that higher test misclassification error rate for female is related with large variance of the female data and the number of diseases associated with dementia

Keywords: Korean Health Insurance Claim Data, Human Disease Network, Bayesian Network, Dementia

INTRODUCTION

The history of mankind has been along with overcoming diseases. The greed which human wanted to live longer, has led to enhancement of medical technologies. And, it lets the human genome project completed. Many statisticians and biologists have used this information from the project, such as revealing the diseases' association. The effort to figure out those rules can be divided into 2 approaches.

The first one is to develop a prediction model. In this method, we can predict the prevalence of target illness, and can figure out other factors related to the target easily. However, there are some drawbacks in this approach. Typically, too many input variables should be regarded, such as human gene information of which the number is around 20,000. Hence, it is hard to estimate all coefficients due to the curse of high dimension. To reduce this shortcoming, shrinkage methods have been developed, including LASSO and ridge regression, which can shrink the number of input variables rapidly rather than an ordinary regression model.

Above all, since the first has limitation, other methods have been remarkable in this place, which is called “Human Disease Network.” A paper from PNAS (Proceedings of the National Academy of Sciences of the USA) at 22th May contained the topic of “Human Disease Network (HDN).” The HDN involves gene information which is concerned with 1,284 disorders corresponded to 1,777 disease genes. Compared to the prediction model, the net has significant advantages in visualization. That is, it helps to analyze overall relationship and clustering. Also from this network, we can discover association that is beyond our stereotypes. Even if this approach has a strong point, the study is quite difficult. The reason is in the data that used, which has massive volume and astronomical cost to obtain.

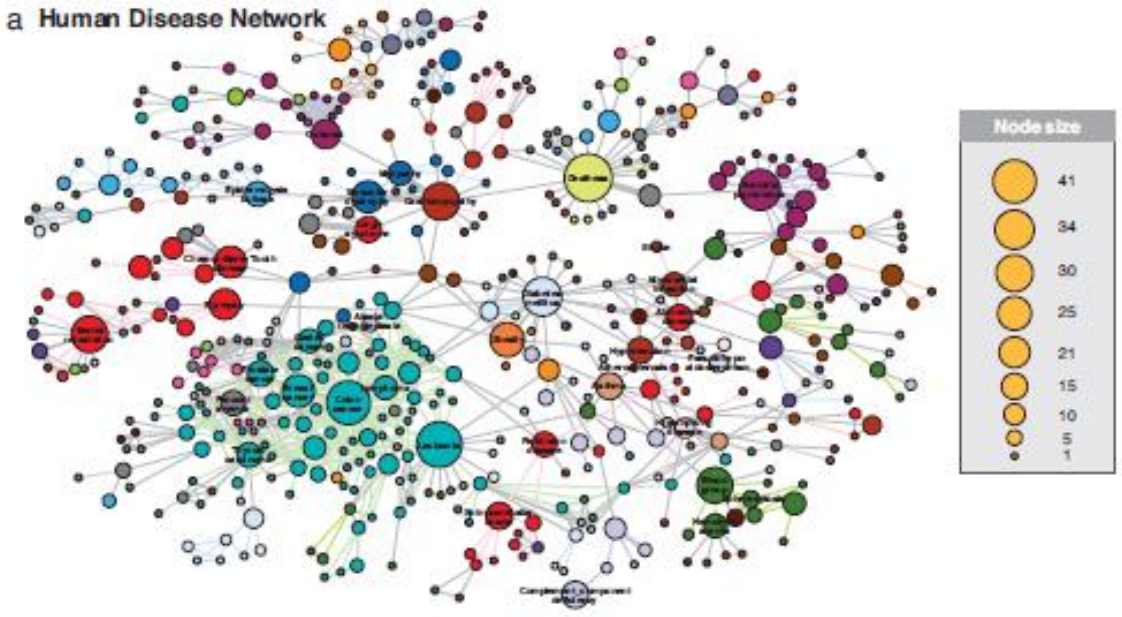


Figure 1. Human disease network in PNAS

To solve this problem, the data from Health Insurance Review and Assessment (HIRA) in Korea has been used, as an alternative of gene data. There are two reasons why I have used the HIRA data set.

1. You can easily obtain data related to one’s disease and trace the patient’s illness history.
2. You can know not only patient’s physical conditions, but also social factors including income and types of health insurance.

The first evidence argues that it can be a substitution of the disease information in PNAS. Of course, some of you think that gene data is more accurate information, so that there is no competitive edge over it. Associations of diseases, however, cannot be concluded only by gene factors, which means social factors also can be major causation of disease in broad[1]. For example, smoking is a well-known trigger of lung cancer. Besides, income can be considered as a factor of some disorders, since it makes difference in health insurance and it derives the large gap of caring health services.

Therefore, what I want to suggest in this paper is that the human disease network with the insurance data which can explain a consequential part of disease. Furthermore, I made the prediction model for dementia with the associations that I had made and other social information.

[1] Austin Bradford Hill. 1965. “The Environment and Disease: Association or Causation?” Proceedings of the Royal Society of Medicine, 58(50): 295-300.

DATA DESCRIPTION AND PREPROCESSING

DATA DESCRIPTION

The data set contains around 1 million patients' records (988,195) for 5 years (2010 ~ 2014). The data set has 2 parts. The first part involves patients' disorders, including main-diseases and sub-diseases. The main-disease is a set of disorders that patients mainly concerned, but sub-disease is a set of incidentally discovered disorders, when a patient has a medical treatment. Besides, only 91 diseases are involved in the data, especially disease that Koreans mainly have concerned.

At the second table, sex, age, insurance and other specific patients' conditions are included. I have used the first table to construct the visualization of the network. This is because that the objective of this study is in the consequential meaning of the data, so that it is more plausible to use the table only involving the disease status of each patient. Through the consequential meaning of the network, I would like to interpret the visualization and the hidden factor of the association of diseases.

DATA EXPLORING

You need to explore the data before analyzing it, because you can have an insight from this step and the analysis can be more precise. In order to scan the data, I have used two methods; frequency analysis and contingency table analysis.

Frequency Analysis

What is the most curious thing if you meet the data set first? You might want to know the number of patients who have had a specific disease. Frequency analysis is the best and the easiest way to show this number. Moreover, this method can show the history of diseases of patients, through filtering patients who have been with the specific illness.

For the frequency analysis, I have generated dummy variables, which are 1 if the patient had record that he or she have ever had the particular disease, otherwise 0. Then, summing up those dummy variables for each 91 diseases, the number of people who have been the illness. A form of the data is similar to a correlation matrix in that the data has symmetric position. Among the results, the bar graph using VA for high blood pressure is following Figure 2.

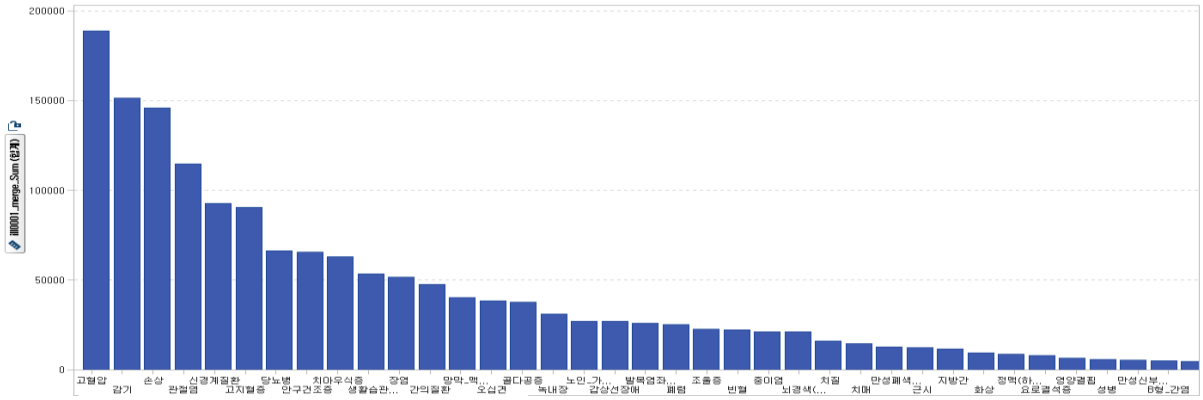


Figure 2. Frequency analysis with high blood pressure

The first disease on the left side of the graph is an index disease, which means that this graph shows patients who have ever had the index disease. So, it is natural that the level of the high blood pressure has the highest level. The graph, as you can see, only shows frequencies of each disease. However, it has a strong advantage as the most direct expression of a disease association. For example, when you see the Figure 2, cold is the highest level except the index. It means that patients who have ever been with high blood pressure tends to be diagnosed as having cold. However, it can mention the only

experience part of outbreak or diagnosis, since the time variable have been ignored in this graph. But, the more important thing is that you can figure out the association between high blood pressure and having a cold, using this simple bar graph.

When analyzing all 91 diseases with this approach, strange but very natural conclusion has appeared. *'Cold', 'Injury', 'Dental Caries'* and other diseases that you can commonly see nearby has been high ranked in most of index diseases. To see more precisely, I have compared the number of patients who was with index disease. And the result was that the number of patients with *'Cold', 'Injury'* and so on was critically larger than other disorders, following Figure 3. That is, association can be shown stronger depending on their frequencies. I have defined those associations as spurious associations, which can hinder you from finding significant associations. And to reduce these spurious associations, sub-diseases have been excluded after this result.

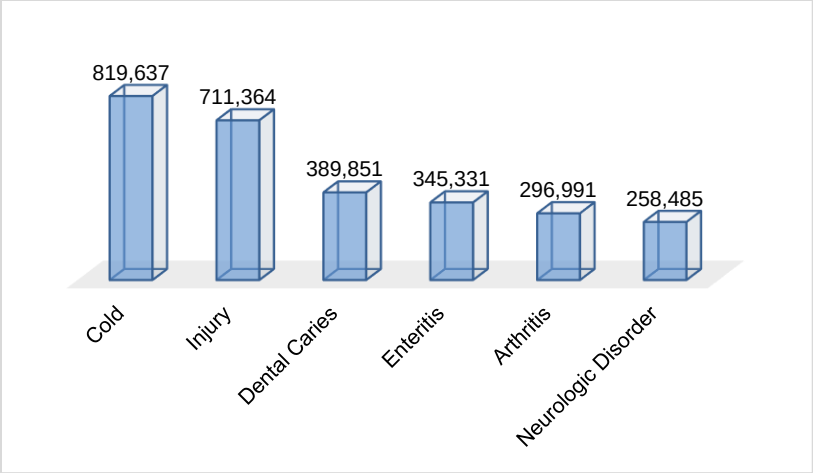


Figure 3. Rank of the number of patients

Nevertheless, we cannot say that all associations that appeared are not significant. There are two things to see the significance. The first one is that disorders related to elder have more significant associations within them. For instance, *'Arthritis', 'Dementia', 'High blood pressure'* and *'Hyperlipidemia'* have been often show in the same bar graph, so that an age variable is relevant to the disease network. And the second is that there are several disorders for female in the same graph, especially diseases related to a uterus. In other words, these female disorders have crucial associations.

Contingency Table Analysis (Crosstabulation)

This analysis is what you can do basically when you are rendering categorical data as the previous one. Especially, given that the data involving dummy variables of 91 disorders, contingency table analysis is meaningful in this case. I have estimated odds ratios of total 4095 pairs of diseases, so that I have cut-off pair of which odds ratio was over 25 and p-value in CMH(Cochran–Mantel–Haenszel) test was less than 0.0001. Moreover, I have described the network using those pairs with SAS® Visual Analytics (VA) to see how the network would be constructed and get an insight from the net applying this preliminary method.

The Figure 4 is the constellation showing for contingency table analysis. Although it is hard to understand instantly, since names of the disorders were written by Korean, each node indicates the name of disease and the width of the link increases, as the estimated odd ratio increases. It has an advantage on the overall description, such as a network, which cannot be attained by the frequency analysis. Remarkable association is the relationship between *'Colorectal cancer'* and *'Hand-foot-mouth disease'* of which the estimated odd ratio is around 293. This seems reasonable, because *Hand-foot-mouth disease* comes from *enterovirus*. Moreover, associations among cancer were also significant. Especially, the odd ratio of 'Oral cavity cancer' and 'Larynx cancer' is 367, and the ratio of *'Precancerosis and Cancer of skin'* is 112.

However, there was limitation. Firstly, there are associations regarding *'Precocious puberty'* with *'Cerebral infarction'* and *'Dementia.'* They are not significant associations in that the data only collected 5 years-diagnoses. In addition, the network showed the association including *'Uterine and Prostate cancer'* which have no intersection at all. Some of you might think that there was a mistake in this result. But, it is a reasonable conclusion, since these exclusive two groups have only dependent proposition within themselves. Although the confidential interval of the odds ratio is tremendously wider than other odds ratio, which indicates ineffectiveness, it is significant in CMH test and in that the interval did not involve 0. Therefore, it is difficult to recognize those spurious association, unless you check all associations in detail.

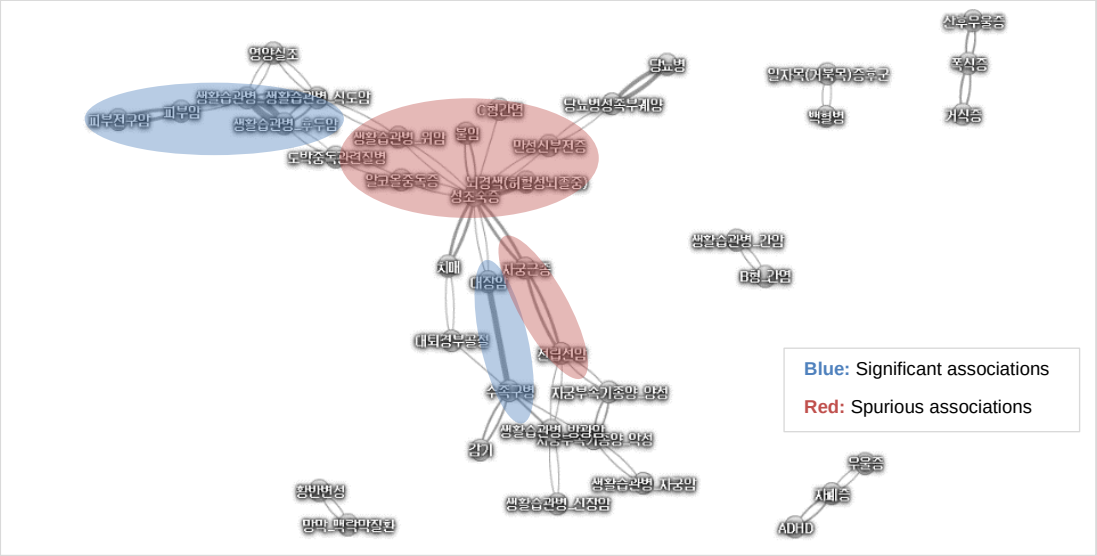


Figure 4. Network from the result of contingency table analysis

DATA PREPROCESSING (DATA RE-BUILDING)

From those results, you can know that it is more accurate when the analysis is done by segmentation of data with sex and age variables rather than using the entire data. Of course, the network including the whole data has own meaning. But reduction of spurious associations is more essential in this case, since the data describes records of diagnoses, which has consequentialism meaning of each patient's history. Thus, you need to re-build the data set.

In order to construct the data set, I have re-sorted dummy variables by the patient's ID in advance. The next is to merge the second table that contains information of patient, including sex, age, address and so on. In this sense, you can prepare the full data set before the segmentation.

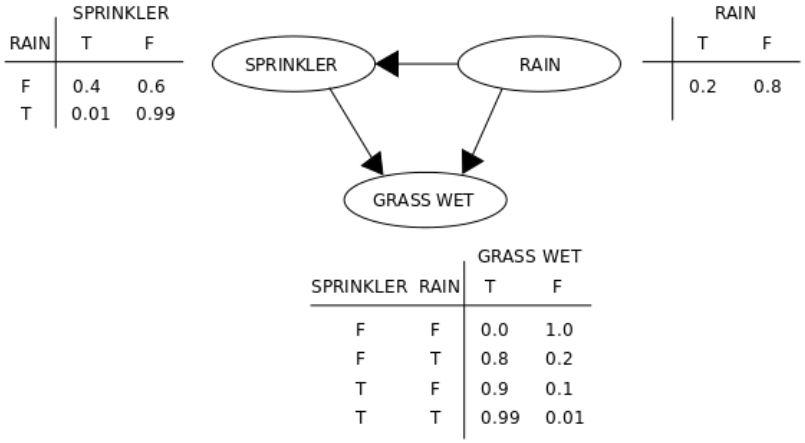
With this data, I divided the data into 8; 4 by age and 2 by sex. Criteria of separation are 20, 40 and 60, so that the data is consist of 0~19, 20~39, 40~59 and over 60. Even though other statistical clustering methods, such as cluster analysis is more appropriate way to fulfill segmentation, continuity of age could be violated in this method. That is, the most essential property of the age variable is broken, so it makes the variable meaningless. Following this process, you can prepare the network modeling.

NETWORK MODELING: BAYESIAN NETWORK

There are several ways to pursue visualization by constructing the statistical network. For example, association rule, logistic regression or even contingency table analysis also can be the solution. But recently, a solution called 'Bayesian Network' has been arisen. Simply speaking, 'Bayesian Network' is a Directed Acyclic Graph (DAG) using a conditional probability. Let's see more detail.

BAYESAIN NETWORK

First of all, you need to know the reason why a conditional probability can describe association among several factors. Suppose that there are only two factors which could cause grass to be wet; either the sprinkler is on or it is raining.[2] And each probability follows Figure 5.



RAIN	SPRINKLER T	SPRINKLER F
F	0.4	0.6
T	0.01	0.99

RAIN T	RAIN F
0.2	0.8

SPRINKLER	RAIN	GRASS WET T	GRASS WET F
F	F	0.0	1.0
F	T	0.8	0.2
T	F	0.9	0.1
T	T	0.99	0.01

Figure 5. Conditional probability of the example

Using the given probability and chain rule of probability, you can easily obtain the joint probability of those events, as follows,

$$\Pr(G, S, R) = \Pr(G|S, R)\Pr(S|R)\Pr(R).$$

With this equation, you can also get each conditional probability that you want. Thus, it is the simple equation.

And the association can be explained by this conditional probability. The conditional probability presumes breakout of a specific event, so that it replaces the sample space with the space of a specific event. Hence, if the conditional probability is sufficiently less, then it can conclude that the association between two events is insignificant. Besides, direction of association can be decided by the basis event, which replaces the sample space.

There are several advantages on this method. First of all, it is intuitively easy to understand the dependencies and joint distribution of probabilities, which is mentioned right before. And the next strong point is that it can deal with the missing value instantly. This is because all variables should be encoded in order to use the Bayesian network, as follows Figure 5. Of course, this model also has the other positive effect, because it is a Bayesian model. For example, it can utilize the prior information of the data. So, you can update the model with the new data. In other word, Bayesian network is a quite suitable method to realize machine learning.

However, the most powerful thing among these advantages is that Bayesian network shows remarkable performance in vague domain. This strength makes the network perform well in nowadays, which is covered from diverse relationship, so that you cannot figure out causes or triggers of events perfectly. In

[2] Wikipedia. "Bayesian Network" Accessed February 15, 2017. https://en.wikipedia.org/wiki/Bayesian_network.

addition, it is more difficult to find out grounds or associations in HIRA's data set which contains only outcomes of diagnoses. But, you can detect latent factors between diseases and grasp the relationship with Bayesian network. There were some studies for patients with 'Heartburn', 'Chest pain' and 'High blood pressure', which are related to a risk factor of 'Heart disease' using Bayesian network.[3]'

On the other hand, Bayesian network has a critical drawback. Like other Bayesian models, its computational cost is more tremendous than other frequentist's methods. Especially, 'Markov Blanket' which is the option in BN classifier node in SAS® Enterprise Miner™ (EM), calculates all possible conditional possibilities, so that the performance of it is time-consuming. In order to solve this problematic situation, you should restrict the structure of the network

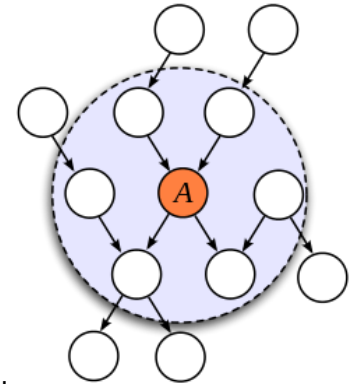


Figure 6. Markov Blanket Structure

In the detailed description, the structure is determined by a parent node and a child node. A parent node indicates a source or an introduction of association. A child node means a result or a conclusion of association. So, the structure of Bayesian network is the result of the linkage rule of parent and child nodes. For example, 'Markov Blanket' structure needs high computational capacity, since it is built with complex rule within its nodes and the illustration is in Figure 6. To prevent high computing cost, I have used two structures; Tree Augmented Naïve Bayes (TAN) and Bayesian Network Augmented Naïve Bayes (BAN).

ANALYSIS

Tree Augmented Naïve Bayes (TAN)

TAN is a sort of Bayesian network model which has more restriction than the simple Naïve Bayes Model. Originally, the naïve Bayes model is the simplest structural network, such that there is no way to know the relationship between child nodes. Especially in SAS® Enterprise Miner™ (EM), when you set a target variable, the target is recognized as a main parent node in EM. Thus, when you perform the simple Naïve Bayes model, it shows only child nodes related to the target node, and linkages of the child nodes are only one related to the target.

Compared to this Naïve Bayes, TAN has more complex structure between nodes as I mentioned. A link from a parent node is connected to a child node. In TAN, however, a child node can be another parent node of other child nodes, which is the most different thing compared to the simple Naïve Bayes. So, TAN shows variables associated with a target variable, and describes relationship between selected variables in the basic level. But, TAN has limitation in that a child node can have only one parent node except a target. In summary, a child node of TAN can receive two parents at most, and one of two should be connected to the target variable. Thus, using the simple rule of the structure, TAN allows you to visualize an overall network rapidly without high computational cost.

[3] Heckerman. D. 1997. "Bayesian networks for data mining" Kluwer Academic Publishers London

Bayesian Network Augmented Naïve Bayes (BAN)

Although TAN has those strength, it is difficult to describe accurate associations due to the simple rule. Besides, in reality, illness cannot be associated with only one disease as TAN. To prevent this misunderstanding, I have used BAN structure. BAN can show more complicated structure in that a child node can receive more than 3 parent nodes including a target unlike TAN. In other words, there is no limit in the number of parent nodes.

Because of this proposition, links in BAN are more complex than TAN. So, it rarely finds the significant association at a glance. However, it has remarkable power, when you observe the network in microscopic way. It also has strength such that it can show each node in causality aspect. In addition, contrasts to Markov Blanket, the target variable only can be a parent node, not a child node. Therefore, it produces less complex network, but computational cost is more efficient than Markov Blanket model.

Figure 7. Structures of TAN and BAN

Application (How to set a target?)

But there is still an important problem, “What variable should be in a target?”. If you set a target variable among the 91 diseases, you will construct 91 disease networks. This process, however, has a lot of disadvantages. Numerous networks make you hard to interpret the network or integrate all network in one network as like as Figure 1 that I have aimed to. Moreover, if you want to eliminate or filter spurious associations through separating data into 8 sets according to sex and age, it is inevitable to face 728 Bayesian networks. Thus, I have thought that another approach or solution should be needed in this case.

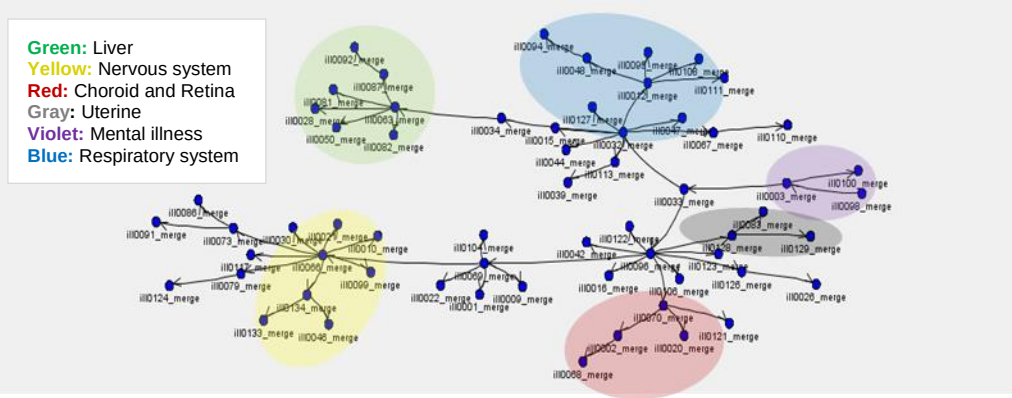
To solve this problem, I have set a target with an age variable. This solution has been derived from the concept of the TAN and BAN of which each child node should be connected to the target variable. Once you set the age variable as a target, illnesses relevant to that age will be selected. Additionally, you can know the associations within disorders in that age easily, since TAN and BAN have other linkage with parent nodes except the target. At the perspective of effectiveness, the data set has already divided into 8 sets by sex and age variables, so that this fact also can be explained by the network. Therefore, the time-varying network can be built when the target is an age variable and input variables are experiences of getting specific disorder which were used in the frequency analysis and the contingency table analysis.

Furthermore, it is obvious true that every child node is linked to the target. Due to this linkage, you are hard to find out the association within disorders. I have erased all linkage with the target and the target node to exhibit the disease’s relationship clearly.

RESULT

Tree Augmented Naïve Bayes (TAN)

From the above strategy, I have constructed TAN first. Instead of showing all constellation, I will only describe TAN for patients over 60 in female. First of all, since links related to the target are eliminated in this constellation, you can see that TAN exhibits only one connection among diseases. This connection makes the plot more similar to tree or branch. In detail, there is a center disease in each branch, and from that disease, illnesses associated with the center disease are stretched out. Thus, you can recognize the cluster of disorders in this age.



Color	Center	Branch		
Green	Disease of liver	Hepatitis B	Fatty liver	Liver cancer
Yellow	Neurological disorders	Cerebral infarction	Cerebral hemorrhage	Dementia
Red	Disorders of choroid and retina	Chronic renal failure	Macular degeneration	
Gray	Myoma uten	Breast cancer	Uterlne tumors	
Violet	Bipolar disorder	Drug intoxication	Post traumatic stress disorder	
Blue	Common cold	Pneumonia	Chronic obstructive pulmonary disease	Lung cancer

Figure 8. An example of constellation of TAN in EM (Female over 60)

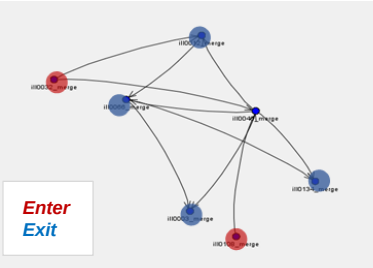
Explanation of each cluster is in Figure 8 in detail. Using this approach, you can check 8 networks of each data set divided by sex and age. Conclusively, each network has different shape and interpretation.

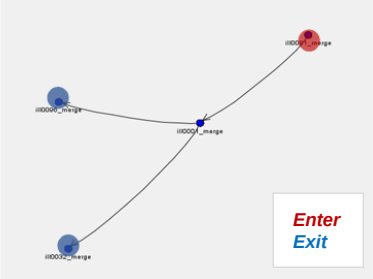
Bayesian Network Augmented Naïve Bayes (BAN)

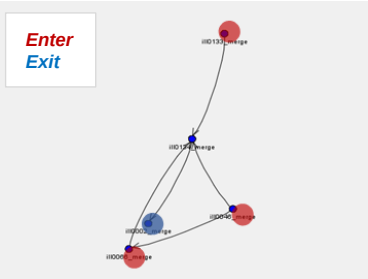
In contrasts, Bayesian Network Augmented Naïve Bayes (BAN) has more complex structure, so that it does not show overall associations well. If you want an overall network description, which is similar to Figure 1, you can apply the approach from TAN. In this study, I wanted to emphasize difference of BAN compared to TAN. To reach this goal, BAN has described a small network related to a specific disorder. And the specific disorder has selected through an index called “Mutual Information.”

Mutual Information is one of the indices in ‘SAS® Enterprise Miner™. This index is a measure for the mutual dependence between two random variables. When it comes to E-Miner, one of those variables set as a target variable of the network, and another variable is one of the input variables in the model. Thus, the most associated disease in each age and sex can be recognized, if you sort diseases in descending order by mutual information. So, I have constructed the network considering 3 diseases with high mutual information. One of the example is shown in Figure 9.

In conclusion, you can grasp micro-association concretely. You can interpret the entrance node to the center disease as an “Enter disease.” In other word, this means disorder that is cause or provoked earlier time-sequentially. Otherwise, called “Exit disease”, which is an output of the center illness. So, these diseases are provoked later in the time measure.







Center	Enter			Exit			
Dementia	*Sepsis*	*Common Cold*	-	*Pneumonia*	*Neurological disorders*	*Bipolar disorder*	*Cerebral Infarction*
High-blood-pressure	*Stomach cancer*	-	-	*Common cold*	*Injury*	-	-
Cerebral Infarction	*Cerebral Hemorrhage*	*Neurological disorder*	*Dementia*	*Diabetes mellitus*	-	-	-

Figure 9 An example of constellation of BAN in EM (Female over 60) from left *Dementia, High-blood-pressure* and *Cerebral infarction* respectively

PREDICTION MODELING

Originally Bayesian network is the core content of this paper. But, I want to figure out whether this network is reliable or not. For this reason, I have constructed the model for dementia through E-miner. There are several reasons to select dementia;

1. Dementia is a representative disorder of the elder in the mutual information approach.
2. There is an increasing trend in the number of dementia patients recently.
3. There are a lot of issues for a measure or criterion for diagnosis of dementia.

To solve these issues, I have developed the model through SEMMA process; Sampling, Exploring, Modify, Modeling and Assessment.

SAMPLING

In the sampling process, you should collect the data and the variable. Variables that I have used are following

1. Sex and Age
2. Main disease code
3. Specific patient code
4. Other diagnosis record
5. Date and Location

EXPLORING

In this exploring process, you should explore data that you had collected in the sampling process. So, I have used the bar graph through 'SAS® Enterprise Guide® (EG)' program. The first result is that the number of female patients are larger than that of male about twice. So, I concluded that there is a significant difference between sex. Of course, in age variable, the elder is more possible to have dementia.

There is also a code for Parkinson's disease in the specific patient code. Moreover, Parkinson's disease is well-known cause for dementia. When it comes to data, there is remarkably significant difference between the Parkinson's and not. However, date and location variables are not significant at all.

MODIFY

In this process, you should define the target of your model first. Through modeling the network, I have already constructed the variable for dementia, so that I used this variable. Besides, in the exploring step, the age variable is significant. I extracted the patients only over 60 to reflect this result.

Secondly, I have re-built the data set to modeling. Hence, I have generated all input variables for main disease codes, specific disease codes, which are mentioned in the sampling. Therefore totally, 335 variables with 278,489 observations were generated. But more important thing is variables that reflect the output of the network. Using mutual information over 0.05, I have made weighted sum for these variables each male and female respectively. Also, using odds ratio from logistic regression that I had made before, another weighted sum variables were generated.

Lastly, you should do data segmentation for modeling. Sex is a well-known criterion for the medical or disease prediction. Besides, in the recent study, 'Delirium' has a significant effect on dementia. Typically, 'Delirium' involves other cognitive deficits, changes in arousal, perceptual deficits, altered sleep-wake cycle, and psychotic features such as hallucinations and delusions. Delirium itself is not a disease, but rather a clinical syndrome. More important thing is that the delirium is often caused by surgery. From this fact, I have made the variable for experience of surgery, and separated the data from this criterion.

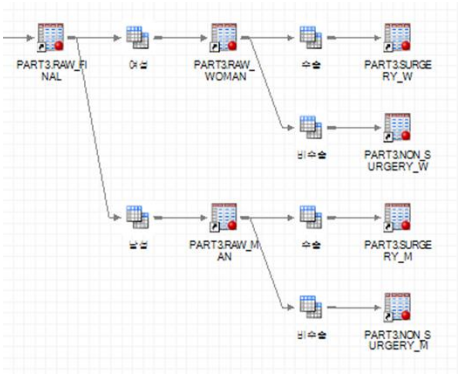


Figure 10. EG nodes for segmentation of data set

MODELING & ASSESSMENT

Firstly, you should sort variables. In this step, you select the useable variables and eliminate the useless one. Through logistic model and odds ratio test, filtering has been done. Therefore, 119 variables were selected for male and 130 variables were chosen for female. Furthermore, it is possible to have collinearity problem, since there are lots of input variables and the same meaning. To reduce this problem, I have filtered some variables with the same meaning. Finally, 104 and 112 variables were selected for male and female respectively.

```
%macro LogitM(var2);
Proc logistic data=part3.raw_man order=data;
Model ill0046_merge=&var2;
Run;
%mend;

%macro LogitW(var2);
Proc logistic data=part3.raw_woman order=data;
Model ill0046_merge = &var2;
Run;
%mend;
```

Output 1. Output from variable selection using logistic regression in macro

```
Proc format;
Value ill 1='Yes' 0='No';
Run;

%macro FreqM(var2);
Proc sort data=par3.raw_man;
By descending ill0046_merge descending &var2;
run;
proc freq data=part3.raw_man order=data;
format ill0046_merge ill. &var2 ill.;
tables ill0046_merge*&var2/cmh1 exact;
run;
%mend;

%macro FreqW(var2);
Proc sort data=par3.raw_woman;
By descending ill0046_merge descending &var2;
run;
proc freq data=part3.raw_woman order=data;
format ill0046_merge ill. &var2 ill.;
tables ill0046_merge*&var2/cmh1 exact;
run;
%mend;
```

Output 2. Output from variable selection using frequency analysis in macro

Model that I had constructed were logistic regression, decision tree and support vector machine (SVM) in EM. And these models applied for each data segmentation and sex. The criteria for selection are test misclassification rate, ROC index and cumulative response rate. I have compared the data segmentation in each three model, and select the best data set in each model to choose the final model. I have already set the training data set 70 % and the test data set 30% before modeling. In this step, I have used the training indices including AIC and the number of variables selected.

Lastly, with the chosen data, I compared considered models by sex separately. Therefore, SVM was selected for male which was divided by the surgery variable and Logistic regression for female which was also divided by the surgery. In case of male patients who have had surgeries, test misclassification error rate is 5.57% and that of the male patients with no surgery before is 3.29%. Meanwhile, the rates for female are 5.45% and 9.73%.

Model	*Segmentation*	*Data*	*Misclassification error rate*	*ROC Index*	*Cumulative response rate*
Logistic Regression	No	RAW_MAN	0.0411	0.871	23.816
Decision Tree	No	RAW_MAN	0.0409	0.826	18.814
SVM	Yes	SURGERY_M	0.0575	0.842	28.945
		NON_SURGERY_M	0.0329	0.871	20.123

Table 1. Assessment models for male

Model	*Segmentation*	*Data*	*Misclassification error rate*	*ROC Index*	*Cumulative response rate*
Logistic Regression	Yes	SURGERY_W	0.0545	0.892	36.0366
		NON_SURGERY_W	0.0973	0.857	46.0817
Decision Tree	No	RAW_WOMAN	0.0675	0.781	32.4668
SVM	Yes	RAW_WOMAN	0.0691	0.871	41.0212

Table 2. Assessment models for female

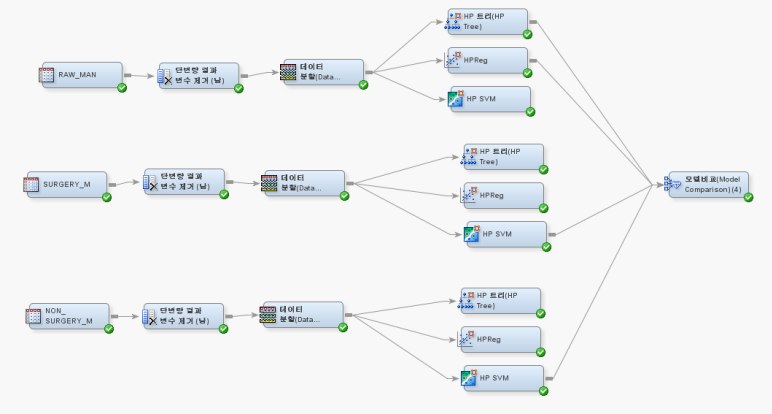


Figure 11. EM nodes for model selection

CONCLUSION

Until now, many analyses have been done. The important thing is that the record data has shown remarkable performance and the network without gene data. What is the reason of this phenomena? There could be several reasons to explain, but in my opinion, the reason is from the power of data in consequential meaning.

When we face an event, we try to understand causes. But, as the world is complicate, it brings numerable causes in one event. So, unrecognizable or even unthinkable reason can be a crucial factor of the result. We called this phenomenon “Butterfly effect” as metaphor. This effect makes the word called reason useless. Moreover, reasons are no more true, which means that it is one of opinions. To illustrates, there are lots of papers which try to proof these opinions objectively.

Of course, the effort to find the reason of events let human develop. So, there is no intend to blame these trials. However, what I have concerned is the stereotype that makes us think that relationship related specific causality is obvious. This thing cannot allow you to see the overall position or association. Therefore, we need Bayesian network using HIRA data, not Figure1.

This study suggests that you can reflect the reality only by analyzing frequency of data for diagnosed patients. Besides, the data has several social or economic details of patients. These contains have the analysis more precise, since gene is the only cause to provoke diseases. There can be other reasons, such as social status or financial state. Hence, it is the wrong conclusion that the result of this paper is coincidence.

In my opinion, further study needs similar approach, but other data including other countries except Korea to see whether analyses show analogous output. And if there is an association that we have not realized before, further study is also needed.

REFERENCES

<Kwang-il Goh, Micheal E. Cusick>. <May 22, 2007>. <The human disease network>. <PNAS> <vol.104 no.21>.

<Austin Bradford Hill>. <1965>. <The Environment and Disease: Association or Causation?> <Proceedings of the Royal Society of Medicine>, <58(50): 295-300>.

<Wikipedia>. <Bayesian Network> <February 15, 2017>. Available at <https://en.wikipedia.org/wiki/Bayesian_network.>

<Heckerman. D>. <1997>. <Bayesian networks for data mining>. <Kluwer Academic Publishers London>

<Trevor Hastie, Robert Tibshirani, Jerome Friedman> <2009> <The Elements of Statistical Learning 2nd edition> <Springer>

<Olivier Pourret, Patrick Naim, Bruce Marcot> <2007> <Bayesian Networks; A practical guide to applications> <Wiley>

<Alzheimer' association> <February 15, 2017> Available at < http://www.alz.org/>

<HIRA> <February 15, 2017> Available at <http://www.silverweb.or.kr/>

CONTACT INFORMATION

Your comments and questions are valued and encouraged. Contact the author at:

<Jinwoo(Kevin) Cho>
<SungKyunKwan University>
<caramelccino@gmail.com>