A Proposal of Test Method for Bayesian Network by Using Self-Organizing Maps

Kunihiro Tada
Mindware Inc.
2-3-10, Shimizu, Okayama, Japan
email:tada@mindware-jp.com

***Abstract*--** In this paper we propose a test method for Bayesian Network by using Self-Organizing Maps (SOM). In order to emulate a Bayesian Network by using a SOM, the generalization capability of the SOM is improved by adding some fictitious data to original training data set, so that the SOM can infer new cases that have been not included in the original training data set. Furthermore, in order to estimate the potential structure of the training data set, potential factor values are estimated from an associated map and they are merged into the original data set. A local regression map from the data set helps to estimate a causal relation between the observed attributes and potential factors.

I. Background and Purpose

As one of usage of Self-Organizing Maps (SOM) [1], it is used to calculate a Conditional Probability like a Bayesian Network (BN). For example, assume that a training data set consists of categorical variables, in which they take the value "1" (truth) or "0" (false). For a case where a certain variable is truth, we can interpret that the mean value of corresponding component in the selected map area on the SOM expression which has the value "1", approximates the probability that the variable takes the value "1".

Even if an attribute has more than 2 states, handling each state as a nominal variable, we can express all states by a SOM. To express the combination of the states of several attributes, it is equivalent to the logic operation AND in the map area with the value of the attributes. Thus, we can approximate the Conditional Probabilities for training data by a SOM.

SOM can usually compress huge data set, but several Mega Byte is necessary to obtain high quality maps that express a nonlinear feature of the data space. In other words, although SOM can preserve much information about data set efficiently, it is not a best choice from the economical point of view, if we want only a rough judgment (i.e., don't want the details). On the other hand, BN needs a smaller memory than SOM, because it deals with only CPT (Conditional Probability Table) to infer suitable judgments.

However, in the case of BN, we have to determine the network structure previously to build the CPT. A network structure can be determined as a Directed Graph from the dependency or independency among the variables. However a user's knowledge is required for determining the direction of the arrows (cause and result) in almost cases. On the other hand, in the case of SOM, we can build the basis for an inference quickly, because we can create SOM maps without using any previous knowledge for the data set. This is an advantage of SOM.

Thus it might be a good idea to build BN for a mixing use, where we firstly use SOM at the training step and finally the inference (recognition) model will be detached from SOM to incorporate to equipment as BN.

Furthermore, when procedural and/or hierarchal judgment is possible, Decision Tree also might be useful. Decision Tree can be interpreted as one of BN that does not have any closed circuits at all.

Now, in this study, we consider a method to support the design process of BN, emulating BN by SOM. The main discussion points are as follows:

1. *To avoid the problem of un-training case by adding* some fictitious data *(improvement of generalization)*.
2. *To extract hidden structure of complex data set, estimating distribution of Underlying Causes by associated map.*
3. *To obtain a strategy for the layering by SOM to build a BN that is layered and adopts Potential Factors (Underlying Causes).*

II. Bayesian Network

In this experiment, Viscovery® Profiler [2] and Predictor [3] developed by Eudaptics software gmbh were used as a SOM tool, and Hugin (GUI) [4] developed by Hugin Eapert as a BN tool. And we consider about Asia.net attached in Hugin as an example.

Fig. 1 shows the Directed Graph for Asia.net. This network has following nodes:

A. Visit Asia?
S. Smoker?
T. Has tuberculosis
L. Has lung cancer
B. Has bronchitis
E. Tuberculosis or cancer
X. Positive X-ray?
D. Dyspnoea? (Breathing difficulty)

Usually, it is used to infer the probability of T, L, B, and E by giving the evidence values of A, S, X and D. It is a significant flexible feature that BN can infer even if there is always not complete information.

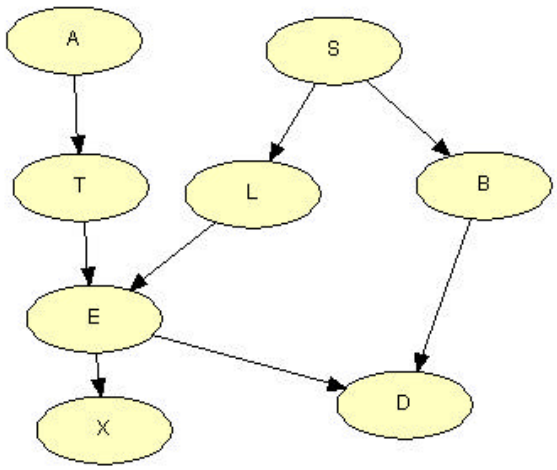


Fig. 1 Directed Graph of Asia.net

The example data named Asia.dat is attached to Hugin, in which the data have 10,000 cases consisting of 8 attribute (corresponds to 8 nodes) which has the states "yes", "no" or "N/A". We can obtain the Directed Graph shown in Fig.1 from Structural Learning by using NPC (Necessary Path Condition) algorithm of Hugin. However, in fact, only a suggestion would be given in a way that user has to complete the Directed Graph by using his/her previous knowledge.

And determining the parameter of CPT by EM (Estimation-Maximization) learning, finally, it will be able to perform the inference. Furthermore, we can create the simulation case by using the built network. This is useful to check the operation of the built BN.

Note however that the percentage of the right answer is not essential, because BN means a judgment about the high indefinite phenomenon. Basically, it would be important to compare the distribution of original data with that of simulation data on the SOM's map.

Strictly speaking, it must be better to compare the probability caused by inference of BN with the probability calculated from database, instead of using only "yes" or "no". However, since a BN can infer about a new case that was not included in the learning data owing to the generalization capability of BN, it is very important to evaluate the inference result about the new case. For this purpose, it needs a criterion for comparison.

Now, let us consider the usage of SOM in order to check quickly the operation of a resultant BN.

III. Problem of Untrained Cases

Even if using SOM, it is unable to express an untrained case by SOM at a usual method. Since SOM summarizes a training data space, it can judge a new case correctly, if such as case is in the training data space. However, if the new case is out of the training data space absolutely, then the inference will be not realized, because the map area by the logic operation AND will be zero. On the practical use, if the probability of untrained cases is very small, then we can expect that the inference result by SOM will be more correct than BN. In this study, since the main purpose is to check the operation of BN, the nodes properties (states) were assigned to a table of orthogonal arrays in order to perform the experiment about also extreme cases that were not included in the training data.

As for SOM, in order to solve the problem of untrained cases, we added the some fictitious data into training data to create SOM's map. Of course, it should not change the distribution of the data too much. In this experiment, we will infer "D" from other node values. 64 fictitious cases were added into 10,000 cases of Asia.dat. That is, for 6 nodes, A, S, T, L, B, and X, they have 2 states (i.e., "yes'' or "no"); if "A/N" is ignored, then all of cases are expressed by 64 cases (2^6).

The values of E and D in the fictitious records are set to "N/A". E depends to L and/or T, whereas D is a target variable in this experiment. Furthermore, for the accurate estimation of the conditional probability, the conditional probability should be the result that divides the total score within the selected map area by the frequency of the training data.

The test of inference was performed 27 times, as assigned to L27 orthogonal arrays with 3 levels, "yes", "no" and "N/A" for the nodes, A, S, T, L, B and X. This experimental result was compared as follows:

SOM1 – non-fictitious cases
SOM2 – including fictitious cases
SOM3 – using the values of SOM2 at only untrained cases on the SOM1
BN – using Bayesian Network

The square root of MSE (Mean Square Error) standardized by the conditional probability by database sorting (DB) is used for this comparison.

Table 1 Comparison of variances for each model

SOM1	SOM2	SOM3	BN
9.446254	20.47332	20.23521	20.41434

Table 1 is the result. Although SOM1 mostly approximates the result from DB, SOM1 and DB have a problem of untrained cases. As a result, SOM3 seems to be the best fitting to BN.

The result of linear regression analysis for BN with SOM3 is:

Linear regression model:

BN = 11.12 + 0.76×SOM3

Adjusted deter. Coefficient	0.868
F-test	171.878
P value	< 0.0001

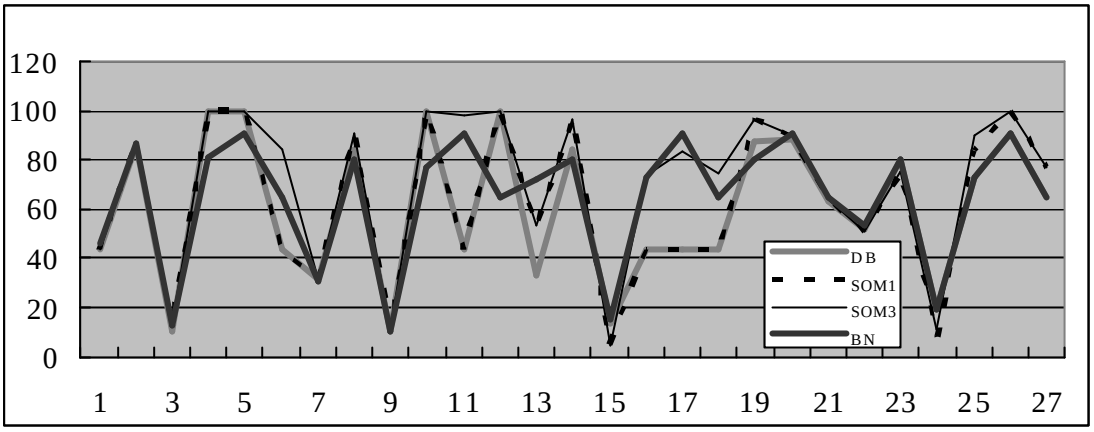


Fig.2 Comparison of the inference result of each model

IV. Extract Potential Factors

As a result, we need not correct this BN model. However, since we can find the hidden structure of the data space by SOM, we will consider the possibility of more advanced model.

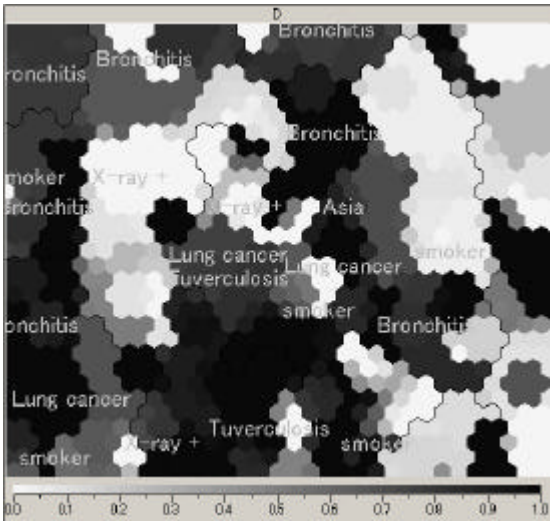
The direct causes of Dyspnoea (difficulty of breathing) must be diseases such as Tuberculosis, Lung cancer, and Bronchitis. It is able to interpret as indefiniteness even if non-Dyspnoea (D=no) despite having some of these diseases. However, there is the case which has D=yes despite T=no, L=no, and B=no. It must be considered that the Dyspnoea was caused by unobserved factors. Now, we create a new factor "V" by selecting map area, and using "Recall" and "Association" facility of SOM. Since this factor was not observed, it must not be an evidence for BN, however it is incorporated into a model by calculating CPT.

Fig.3 is the comparison of the map for the original data and the case that adopted potential factor "V", where "D" was not contributed to the map ordering. It is found that the potential factor "V" contributed significantly to the map. The comparison of the residual by using SOM local regression model of Viscovery® Predictor is shown as follows.

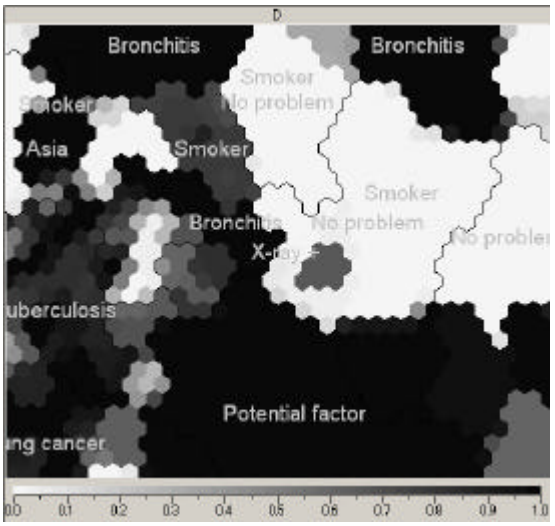
The model without potential factor:	
Linear explained variance	*45.4 %*
Nonlinear explained variance	*7.8%*
Nonlinear residual	*46.8 %*
The model with potential factor:	
Linear explained variance	*64.7 %*
Nonlinear explained variance	*28.3 %*
Nonlinear residual	*7.0 %*

Note that it does not indicate the performance as a predictive model, because the potential factor "V" was not observed. However, the potential factor gave more explanation about the variance. The purpose of this potential factor is to extract a hidden structure of data space.

In general, it further must be helpful for interpreting the potential factor to create a linear regression model and a SOM local regression model that sets the potential factor "V" as a target variable. However, in this research, it was omitted because the linear-regression model did not indicate any significant result.



(a) Without potential factor "V"



(b) With potential factor "V"

Fig. 3 The local regression map targeted "D"

The left bottom areas, which are labeled "Asia", "Tuberculosis", and "Lung cancer" in the map of Fig.3 (b), correspond only to 9.5% of the whole cases. Thus, 90.5% of people did not visit Asia: they do not have not only Tuberculosis but also Lung cancer; in other words, if he/her

does not have not only Bronchitis but also potential factor, then he/she is not Dyspnoea. However, from this data we cannot know the answer about "what is the potential factor?" However, we found that the risk due to the potential factor is only 7.9%.

V. HIERARCHAL MODEL

Thus, 90.5% of whole people are able to know the accurate probability by a Decision Tree in Fig.4 rather than BN. For the case of A=no, T=no, L=no, and B=no, the accurate probability by the Decision Tree was 12.1% but BN was 10%; for the case of B=yes, the Decision Tree was 75.4% but BN was 80%.

On the other hand, since the area of "Asia", "Tuberculosis" and "Lung cancer" is different from that of the right area on this map, we extract the corresponding data of this area and create a more detailed model.

Fig.5 is the resultant map. We found visually that D correlates with E and B. Also the left bottom area on this map corresponds to the potential factor "V".

Hence, if the variance of the areas $B < 0.5$ and $D \geq 0.5$ was explained, the structure of the map would be shown more clearly. We assume the potential factor "W" again. Fig.6 is the corresponding map. The risk due to the potential factor is about 21%.

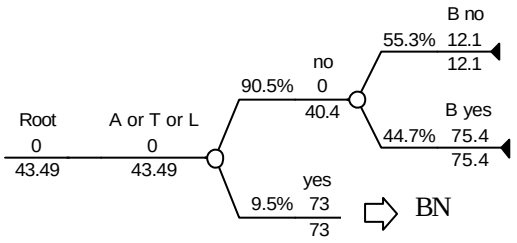


Fig.4 Decision Tree

The model of Fi.5:

Linear explained variance	*32.6%*
Nonlinear explained variance	*6.2 %*
Nonlinear residual	*61.2 %*

The model of Fig6:

Linear explained variance	*62.6%*
Nonlinear explained variance	*23.0 %*
Nonlinear residual	*14.4 %*

Fig. 7 (a) and (b) show F-significance on the maps of Fig.5 and Fig.6. Black area on the map indicates the area of significant model (a) and highly significant model (b).

However, these SOM local regression models are not for prediction but for interpreting the structure of data space. As for the linear regression model which considers W as a target variable, the Adjusted Determination Coefficient was 0.476, where each Regression Coefficient is as follows:

T：0.18276012	*B：0.30620848*
L：0.11317900	*X：0.05806980*
S：-0.08879527	*V：-0.66618502*
A：-0.19579882	*Intercept: 0.38213270*

In the case of SOM local regression model, 15.8% of variance was explained additionally and the residual was 36.6%.

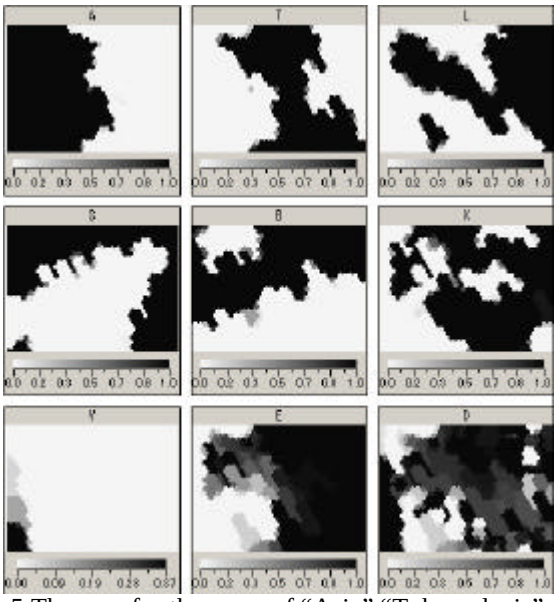


Fig.5 The map for the cases of "Asia" "Tuberculosis" and "Lung cancer"

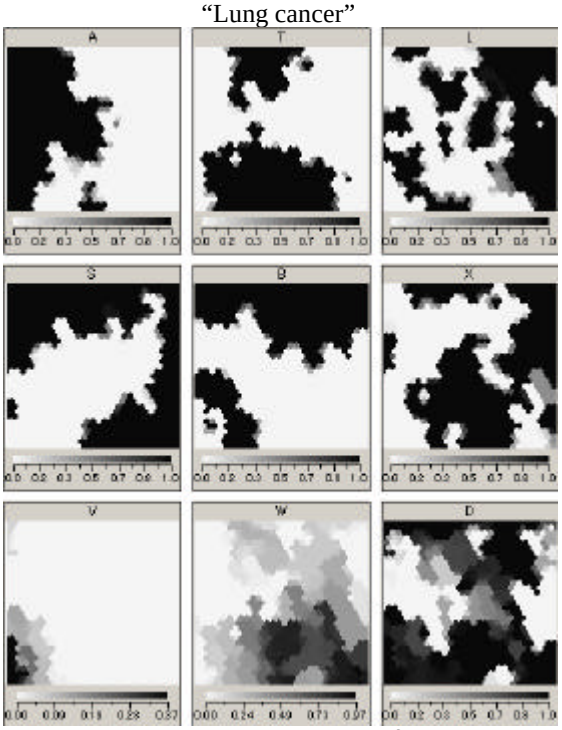
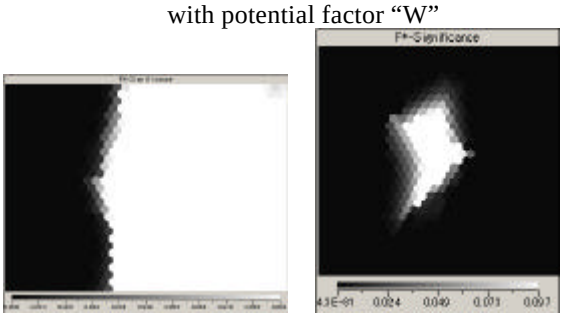


Fig.6 The regression map for D with potential factor "W"



(a) (b)

Fig.7 F-Significance

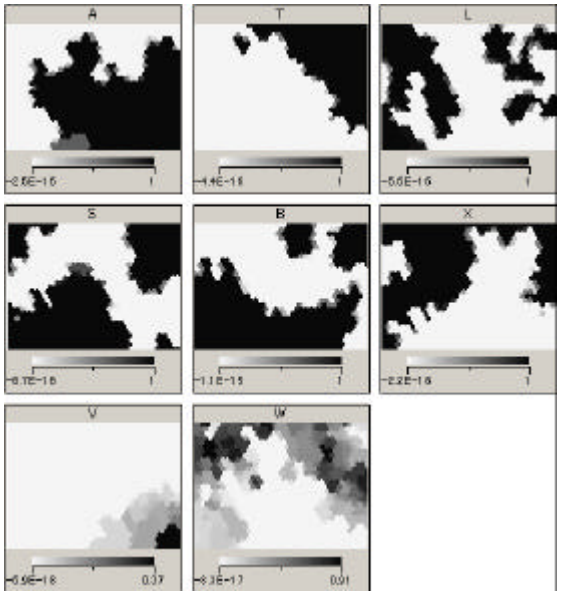


Fig.8 The regression map for the potential factor "W"

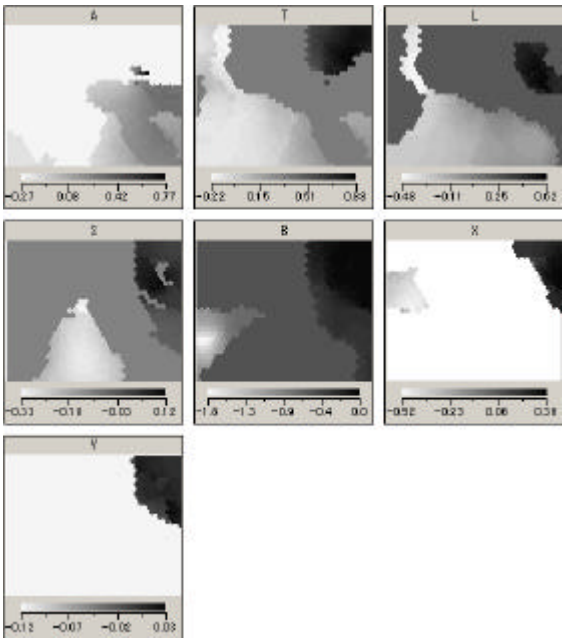


Fig.9 The local regression coefficients for the potential factor "W"

From the above-mentioned result, the sub model for the sub area of "Asia", "Tuberculosis" and "Lung cancer" was created. Fig.10 is the corresponding Directed Graph.

The linear regression model:

BN = 12.20 + 0.69×SOM

Adjusted deter. Coefficient	0.468
F-test	23.864
P value	< 0.0001

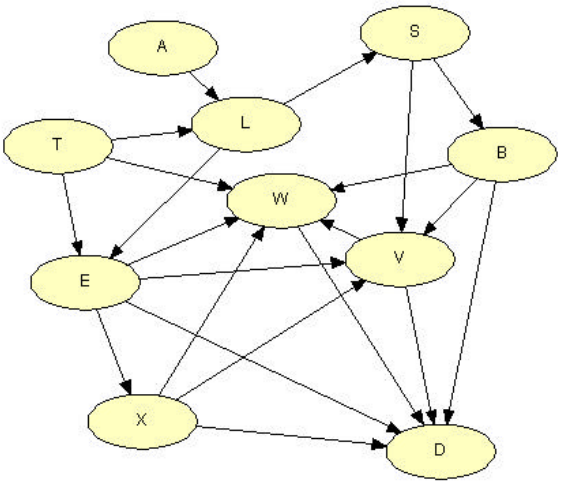


Fig.10 The Directed Graph for the cases of "Asia" "Tuberculosis" and "Lung cancer"

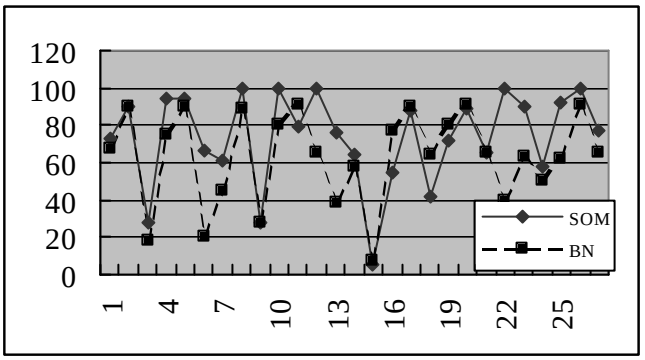


Fig.11 Comparison graph for the cases of "Asia" "Tuberculosis" and "Lung cancer"

VI. Conclusion

The adding of some fictitious data was found to be effective for improving the generalization facility of SOM. It must be applicable to other applications, not only to emulate BN.

We could discover a potential factor; however it has a different meaning from the statistical potential factor. Thus it means discovering a new "concept", and suggests that the pattern recognition or association like a SOM is necessary as basis for an inference. In the future, we'd like to consider a combination of SOM with other statistical techniques that use potential factors.

References

[1]T. Kohonen. *Self-Organizing Maps*, Springer-Verlag Tokyo, Tokyo, 1996

[2]*Viscovery® Profiler version1.2 User's manual*, Eudaptics softwer gmbh, Vienna, 2001.

[3]*Viscovery® Predictor version1.1 User's manual*, Eudaptics softwer gmbh, Vienna, 2002.

[4]*Hugin GUI Help*, Hugin Expert A/S, Aalborg, 2003.