

## DATA APPROXIMATION FOR BAYESIAN NETWORK MODELLING

Shaoning Pang

Knowledge Engineering and Discovery Research Institute  
Auckland University of Technology, Auckland, New Zealand  
[spang@aut.ac.nz](mailto:spang@aut.ac.nz)

### ABSTRACT

Inspired by results of previous works on data approximation for BN, probability norm minimizing (PNM), which reflects the proximity of both variable values and frequencies in probability space, is proposed as the criterion for BN intra-variable learning. For the solution of this objective function, we employ a kernel-based estimation, and design a kernel-based iterative algorithm for BN learning. Our data approximation simulation on UCI data CLOUD has proved the convergence of PNM and its superior approximation accuracy to the typical histogram method. Furthermore, we have integrated our proposed PNM in BN learning on both real data and UCI data. The results demonstrate that PNM has endowed the learned BN with better performance of fitting data for its remarkable classification accuracy promotion than previous work (EL-Matouat, F. et. al. 2000). More significantly, it can help the BN to disclose the correlation of the dataset in multiple resolutions, and provide a better fit to prior knowledge.

**Keywords:** Bayesian network, Minimum Norm, Data Approximation, Intra-variable learning, Inter-variable Learning, Kernel Based Estimation

### 1. INTRODUCTION

Because of the prevailing property in encoding uncertain information, Bayesian network techniques have been widely used and have been shown to be remarkably effective in various data-analysis applications (Campos De et. al. 2000, Kirsch, H. et. al. 1994, Suojanen, M. et. al. 2001). Especially in expert systems, it can provide plausible probability estimation on such diagnosis problems as "Given a certain assumption, what is the prediction result?"

A Bayesian network is a graphical model that discloses the relationships among variables of data sets with Bayesian probability, in which the structure of the graph model illustrate the association of objects, and probability is used to evaluate the relations quantitatively. Further, both the structure and the probability distribution are learned from the data set.

Learning Bayesian network from data can be eventually explained as a searching procedure in a network hypothesis space. It is to identify a network structure with high scores by some criteria. At this point, many criteria such as maximum likelihood, minimum cross entropy and some extended likelihood like AIC and BIC etc. (Poland R.B. et. al. 1994, Sclove S.L. 1994), are popularly used to measure how much the learned BN model fits the data and prior knowledge (Herskovits E.H. et. Al. 1995, Geiger D. 1992). Nevertheless, we have noticed that most criteria-based BN learning methods are exclusively inter-variable, in which all the optimization criteria are defined within the variables, but regardless of the necessary optimization inside the individual variable. Fortunately, in the work of (EL-Matouat, F. et. al. 2000), El-Matouat et. proved that the optimization of the individual variable is advantageous and necessary to BN learning.

For data approximation of variable, many previous researches (EL-Matouat, F. et. al. 2000, Viswanath P. et. al. 1996) indicated that frequencies and values are two key measurements for data approximation. El-Matouat et. al firstly applied the intra-variable optimization to model the belief network classifiers, and proposed an optimal histogram-building scheme, which induces a sampling of continuous variables existing in the network. The problem with their method is that it is fundamentally frequency-

based, which eventually cannot exclude the inherent drawback of histogram, also is short of a mechanism to deal with prior knowledge behind the data.

This paper addresses the problem of optimal data approximation for BN. A new criterion of probability minimum norm is proposed to realize such optimization inside each variable of the dataset. The experimental results either on intra-variable simulation or on BN learning have demonstrated the effectiveness of our proposed methodology. The organization of this paper is as follows: Section 2 introduces some basic knowledge of Bayesian network modeling. Section 3 discusses previous data approximation methods for BN. In Section 4, we propose the PNM data approximation for BN modeling and give out the learning algorithm. In Section 5, after the comparing simulation of PNM for convergence and accuracy test, we have implemented the proposed PNM to learn BN from a UCI database and a real database respectively. Finally, we present our conclusions and directions for future work.

## 2. BAYESIAN NETWORK MODELLING

To build a Bayesian network, one needs to specify two things to describe a BN: the topology (structure) and the parameters of each node (conditional probability distribution CPD). This is an NP-hard problem, since the number of DAG's is super-exponent of the number of variable  $n$ . If there are  $k$  variables related to the model, then  $k(k-1)/2$  different undirected arc one can place on the network. It follows that the size of network hypothesis space will be  $2^{k(k-1)/2}$ . Thus, researchers have used heuristic searching algorithms, such as greedy search, EM algorithm, best-first search, and Monte-Carlo methods etc. In our network learning experiments of this paper, a two-pass greedy search algorithm, as below, is employed.

Suppose a network structure  $S$  is known and all the variables are observable. We denote  $S^h$  as the hypothesis that the joint probability can be factored according to  $S$ ,  $V$  as the set of case variables in the network  $S$ , and  $\Delta C$  as the change of the criterion score defined for network learning, which we will discuss the detail in next section.

Step 1: Choose a network hypothesis  $S^h$ , evaluate  $\Delta C$  for all the arcs in  $S^h$ , make the change of arc in two passes for which  $\Delta C$  is maximum, provided it is positive. In the first pass, arcs were added until the model score did not improve. In the second pass, arcs were deleted until the model score did not improve.

Step 2: Repeat step 1 until there are no arc changes with a positive value for  $\Delta C$ .

An important measurement for BN learning that are often used for model selection is the log marginal likelihood  $p(D | S^h, \vartheta_s)$  (Geiger D. 1992, Kullback S., 1959), where  $\vartheta_s$  is the vector of network parameters, and  $S^h$  denotes the network hypothesis,  $D$  is the dataset with the assumption of being without the missing data.

If we assume  $D$  with unrestricted multinomial distributions, parameter independence and without missing data, each parameter vector  $\vartheta_{ij}$  is updated independently. For every  $i$  and  $j$ , the marginal likelihood of the data is just the product of the marginal likelihoods for each  $i-j$  pair

$$p(C | \xi) = \frac{\Gamma(\alpha)}{\Gamma(\alpha + N)} \prod_{k=1}^r \frac{\Gamma(\alpha_k + N_k)}{\Gamma(\alpha_k)} \quad (1)$$

$$p(D | S^h) = \prod_{i=1}^n \prod_{j=1}^{q_i} \frac{\Gamma(\alpha_{ij})}{\Gamma(\alpha_{ij} + N_{ij})} \cdot \prod_{k=1}^{r_i} \frac{\Gamma(\alpha_{ijk} + N_{ijk})}{\Gamma(\alpha_{ijk})} \quad (2)$$

and the well known maximum likelihood approach of learning Bayesian network is:

In network structure hypothesis space, to find a network structure  $S^h$  whose maximum likelihood over parameter  $\vartheta_s$  is the largest

$$\underset{\vartheta^{S^h}}{\text{Max}} (p(D | S^h)) \quad (3)$$

Based on the above Bayesian method, a number of extensions to the maximum likelihood approach have been worked out in one form or another (Cooper G. F, et. al. 1992, Cowell R.G., et. al. 1999) to overcome some problems like over-fitting, etc. These approaches replace the sample likelihood by a modified score

that is to be maximized, such as the penalized likelihood, Akaike information criteria (AIC), the Bayesian information criteria (BIC), and others (Poland R.B, 1994 , Sclove S.L. 1994 , Geiger D. 1992). In our experiment, we use only the maximum likelihood criteria in our network structure searching.

### 3. THE PROPOSED METHODOLOGY

For the problem of individual variable approximation, the procedure can be classically described as below:

Let be  $X = A_1, A_2, \dots, A_p$  an initial partition of an unknown distribution  $\lambda$ , apply a certain data approximation model  $f(X)$  on this initial partition, we can have a model optimal distribution  $\lambda_c$  with the sub-partition of  $X_c = B_1, B_2, \dots, B_c$  and  $c \leq p$ . Then the goal of any above procedure is to approximate well the entire data distribution  $X$ . Thus, the problem concerns how to fix an optimal measurement.

By the results of most previous works on data approximation (Ioannidis Y, et. al. 1995, Haas P.J., 1995, Korn, F et. al. 1999), “value” and “frequency” approximation optimization are consistently emphasized as the two basic elements to fit the data. That is, only when approximate data in both “value” and “frequency” view of data distribution, then such data approximation can be counted as the best fitting. However, for Bayesian network, probability calculation is basic, either BN structure or its parameters computing are typically existed in probability space. Obviously, to define “value” and “frequency” criterions using probability is necessary and certainly will be beneficial to the working of BN, which has been proved in (EL-Matouat, F. et. al. 2000).

#### 3.1 Probability Norm Minimum (PNM)

To overcome the shortcomings of above Intra-variable learning methods, we propose a new criterion, in which the probability norm is defined to teach the Bayesian network intra-variable learning. It is depicted below:

##### Case of Single Variable:

Given domain data set  $D = \{X_1, \dots, X_n\}$ , for each variable  $X_i$  of  $D$ , we define the norm below to identify the approximation error in both probability frequency space and value space.

$$\|U_i\|^2 = \int_{-\infty}^{\infty} (p(x_i)f(x_i))^2 - (p(x_i)x_i)^2 dx_i \tag{4}$$

and sum the norm difference of all the variables in  $D$  as below, then  $\|U\|^2$  is the identification of approximation error.

$$\|U\|^2 = \sum_{i=1}^n \|U_i\|^2, \text{ where } \sum_{i=1}^n p(x_i) = 1 \tag{5}$$

The approximation searching criteria for  $D$  is

$$\text{Min} : \int_{-\infty}^{\infty} (p(x_i)f(x_i))^2 - (p(x_i)x_i)^2 dx_i \text{ with } i = 1, \dots, n \tag{6}$$

##### Case of Multi-variable:

Based on above single variable probability norm definition, it is easy to extend the above equations to the case of bi-variable with the conditional probability as below

$$\|U_{ij}\|^2 = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} (p(x_i | x_j)f(x_i))^2 - (p(x_j | x_i)x_j)^2 dx_i dx_j \tag{7}$$

$$\|U\|^2 = \sum_{i=1}^n \sum_{j=1}^n \|U_{ij}\|^2 \tag{8}$$

However, the case of bi-variable usually has more computational complexity. It can be used to some special database analysis, in which the independence assumption between variables is not valid any more, such as some scientific experiment datasets.

**Kernel-based Solutions:**

For the solution of Eq.(9), Kernel based estimations have been proved in mathematics and also have been widely used as a simulation for such optimization. It is one of the most popular methods in statistics. The basic idea is simple: for each data point  $X_i$ , a kernel (e.g., a Gaussian) centered about  $X_i$  is summed.

**Definition:**

Given data points  $X_1, X_2, \dots, X_n$ , a kernel estimation function  $f$  is constructed as follows

$$f(x, n) = \frac{1}{n} \sum_{i=1}^n Kh(x - X_i) \quad (9)$$

where  $n$  involves with the level  $L$  of approximation,  $Kh(x)$  is usually a unimodel, symmetric and bounded density function depending on the bandwidth  $h$ , typically it can be one of the functions as below:

**Gaussia:**

$$Kh(x - X_i) = \frac{1}{\sqrt{2\pi}h} e^{-\frac{(x-X_i)^2}{2h^2}} \quad (10)$$

**Polynomial of degree  $d$ :**

$$Kh(x - X_i) = (1 + xX_i)^d \quad (11)$$

**Gaussian radial basis function:**

$$Kh(x - X_i) = e^{-\|x-X_i\|^2} \quad (12)$$

**Cubic splines:**

Given knots  $a \leq x_i \leq b$  with value  $u_i$  at each knot, then a cubic spline is an interval function on  $[a, b]$  as:

$$f(x) = a_i + b(x - x_i) + c_i(x - x_i)^2 + d_i(x - x_i)^3 \quad \text{if } x \in [x_i, x_{i+1}] \quad (13)$$

**Sngnoid:**

$$f(x) = \begin{cases} x_i & \text{if } x \geq x_i \\ x_{i+1} & \text{if } x < x_{i+1} \end{cases} \quad \text{where } x \in [x_i, x_{i+1}] \quad (14)$$

**3.2 Algorithm for PNM Data Approximation**

According to the above approximating model, the process of data approximation can be treated as a kernel-based approximating procedure with the criterion of Eq. (6). To ensure the globally optimal approximation, we act the criterion Eq. (6) iteratively on the variable. That is, first minimize the norm Eq.(6) with a kernel estimation in the whole the definition range of the input variable, and find the initial partition point, with which we can divide the variable into two sub-ranges. Then perform the same operation on each of sub-ranges iteratively, until it can't be divided or the iteration level decreases to zero. The Algorithm is depicted below:

**Algorithm 1:** Kernel-based Iterative Approximating (for single variable)

**Inputs:** Approximation variable ( $x$ ), Kernel choice (*kernel*), Iteration layers (*level*), minimum norm threshold ( $\xi$ ).

**Output:** Range Selectivity of input variable (*ooth*, *oodth*)

**Procedure** double \* **Intra\_variable\_learning**( $x, L, \text{kernel}, \xi$ );

Static int C out\_c[]; /\* Define static global variable to record the range selectivity \*/

$L = \text{level}; C = 1;$  /\* Variable initialization \*/

Compute  $f_c(x, L)$  by Eq. (9)

Compute initial PNM approximation  $\|U_0\|^2$  by Eq. (4)

**While** ( $C \leq \text{length}(x) \parallel \|U_c\|^2 < \xi \|U_0\|^2$ ) {

$i = 1;$

**While** ( $x(C+i) \leq x(C)$ )  $i++;$

$C = C + i;$

Update  $f_c(x, L)$

```

Update PNM approximation  $\|U_c\|^2$ 
If ( $\|U_c\|^2 < \|U_{c-1}\|^2$ ) {
     $\|U_{c-1}\|^2 = \|U_c\|^2$ ;
    Note the location  $C$ ;
}
L=L-1;          /* Iteration level counting*/
If (L != 0) {
     $x_{front} = x(1), \dots, x(C)$     /* Organizing the data for iterative searching*/
     $x_{behind} = x(C+1), \dots, x(Length(x))$ 
    If (length( $x_{front}$ ) != 0) Intra_variable_learning( $x_{front}$ , L, kernel,  $\xi$ );
    If (length( $x_{behind}$ ) != 0) Intra_variable_learning( $x_{behind}$ , L, kernel,  $\xi$ );
    L = L - 1;
    Store the searching result in global memory;
    return Null; }
else
    return out_c;

```

#### 4 EXPERIMENTS AND DISCUSSIONS

##### 4.1 Property Test of PNM

Figure 1 is the illustration of PNM approximating in different levels, the results indicate that the PNM approximation error is consistently decreasing with the increment of *level*. Hence, it can be believed that the above data approximation with the proposed PNM criterion can tend to have minimum PNM error.

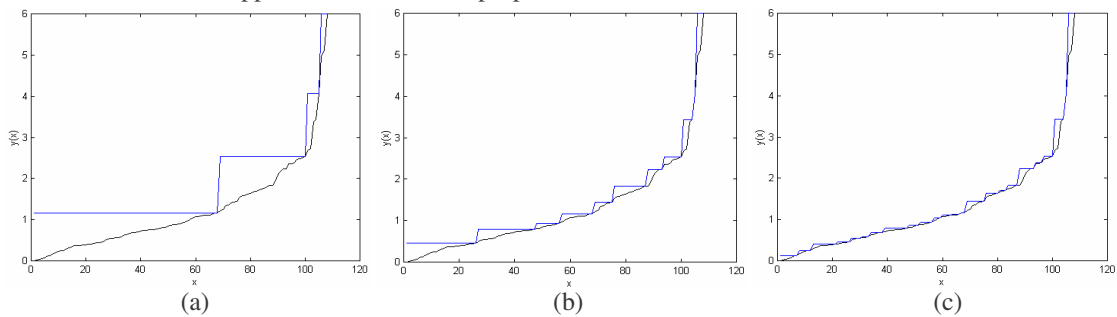


Figure 1. PNM approximating at different levels.

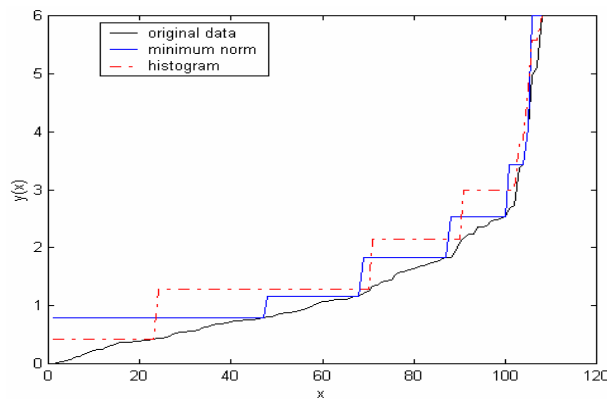


Figure2. The Contrast of approximation  $x$ , of Cloud with histogram and PNM

Next, in order to compare PNM with other data approximation method, with the same resolution, we have partitioned the attribute 2 in the Cloud dataset using a typical histogram method and PNM respectively. The result is as Figure 2, from which we can see that PNM is more accurate to approximate than the histogram, So it presents us with a more detailed estimation of the data distribution.

**Heart Disease Database**

Follow the experiment of (EL-Matouat, F. et. al. 2000), we also have implemented our proposed methodology on the same heart disease database given by the UCI machine-learning group. This database contains 13 attributes to record the results of various medical tests carried out on different patients, which are indexed from A to M in Table 1. The purpose of this dataset is to predict the presence or absence of heart disease. We perform the Algorithm 1 only on the attributes appeared with an asterisk in Table 1, and train the Naïve-Bayes network [6] to predict disease, the resolution here we take is Resolution 2 (*level=3*) listed in Table 2

Table 1 Contrast results of different approximation methods on the Heart database

%	Diagnosis Error	
	Absence	Presence
<b>Equal range</b>	28.6	35.6
<b>Histogram</b>	15.8	31.3
<b>Optimal Histogram</b>	14.1	13.3
<b>PNM</b>	12.0	8.7

The final results are presented in Table 1, our proposed method obtains almost 90% good classification by taking into account the costs with PNM towards the best result 86.0% reported in ((EL-Matouat, F. et. al. 2000).

Table 2. Resolutions of PNM inter-variable learning

Resolution	Number of states												
	A*	B	C	D*	E*	F	G	H*	I	J*	K	L*	M
<b>Resolution1</b>	7	2	4	7	3	2	3	7	2	3	3	4	3
<b>Resolution2</b>	14	2	4	14	7	2	3	15	2	7	3	4	3
<b>Resolution3</b>	25	2	4	22	14	2	3	28	2	14	3	7	3

In order to test the influence of the PNM learning resolution (parameter *level*) on the construction of Bayesian network, we employ three resolutions in Table 2 with *level* equal to 2, 3, 5, respectively, on the Bayesian network learning.

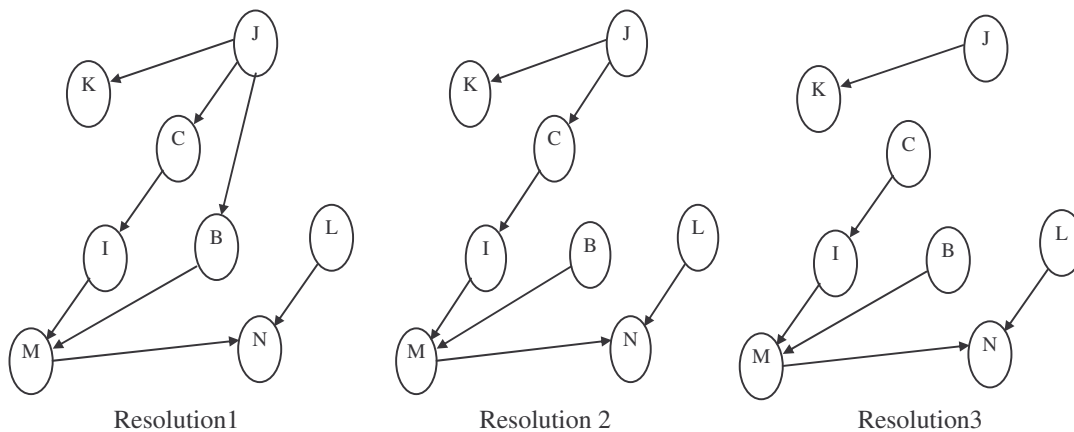


Figure 3. The resulting Bayesian network structure with PNM in the resolutions of Table 2.

Figure 3 is the resulting BN structures; from which we observe that different resolution we used in PNM learning will lead to different BN structure. Compare the cross validation of three different cases, Resolution 2 gets 87.91%, which is superior to 85.57% and 87.24% of the other two cases. That indicates a proper choice of approximation resolution is advantageous for BN to fit the data.

**Telecom Fraud Database**

A second database is a real database on mobile phone payment. It includes 15 tables in our data mart, which records 53,696 user one year’s payment action for fraud detection. Among these 15 tables, three core tables for the topic are IBS\_USERINFO, IBS\_USRBILL, and IBS\_USERPHONE and all others are the additional tables. At the feature extraction and selection stage, SAS DATA PROC is used to calculate 7 numerical attributes to measure the customer payment performance; they are Delay\_Time, Delay\_Total, Delay\_Num, Fee\_Total, Fee\_Freq, Fee\_StdFreq, Delay\_Rate, then a PROC CLUSTER is used to evaluate the credit grade by the payment performance. Finally, a PROC SQL, which contains the following SQL sentence, is taken to combine the IBS\_USERINFO and IBS\_BILL to IBS\_USERPERFORMANCE.

```
select a.CreditDegree, d.*
from Fraud.ibs_Credit as a, Fraud.ibs_Bill as b, Fraud.ibs_Usrphone as c, Fraud.ibs_Usrinfo as
d
where a.billphone_id=b.billphone_id and
b.account=c.account and
c.usrinfo_id=d.usrinfo_id
into Fraud.ibs_Userperformance;
```

Table IBS\_USERPERFORMANCE has 53,696 records and 8 attributes, which contains one class attribute, Fraud\_Id, and above 7 numerical attributes, and which, in this paper, are indexed as 8 capital characters from A to H.

In this experiment, we apply two data approximation methods to BN variable learning, one is a typical histogram, and the other is our proposed PNM. With the two-pass greedy search algorithm that we introduced above, Learn BN from dataset IBS\_USERPERFORMANCE under the rule of Eq.(2), the prediction results in Table3 prove that PNM shows remarkable improvements towards the typical histogram method.

Table 3 Contrast results of different approximation methods on the Fraud database

%	Prediction Error	
	Fraud	Non-Fraud
<b>Histogram</b>	13.6	10.6
<b>PNM</b>	4.2	2.7

The BN structures we obtain from those two methods are shown as (a) and (b) respectively in Figure 4. And the superior fraud prediction accuracy demonstrates that structure (b) is better fitting dataset IBS\_USERPERFORMANCE than structure (a).

Table 4. Number of states for the variables of the Fraud database

Resolution	Number of states							
	A	B	C	D	E	F	G	H
<b>Resolution1</b>	4	8	8	10	8	4	4	8

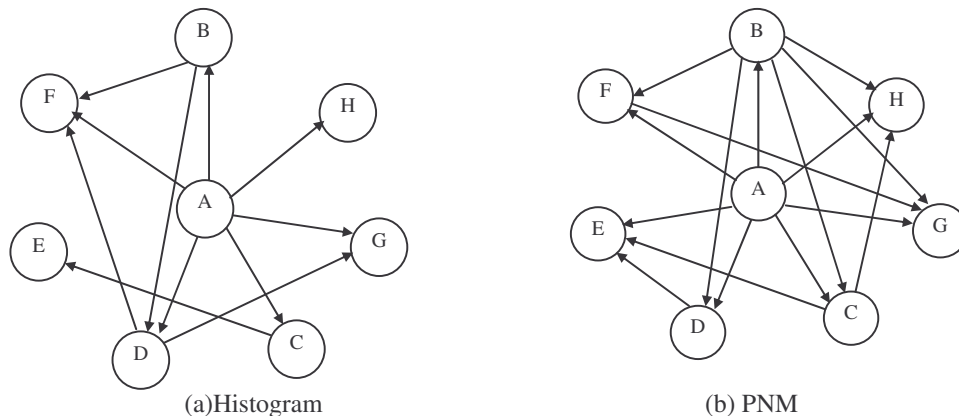


Figure 4. The Bayesian network structure learned from fraud database with PNM and Histogram

Additionally, in structure (b) of Figure 4, some important correlations, such as  $A \rightarrow E$  (Fraud\_Id  $\rightarrow$  Fee\_Total),  $C \rightarrow H$  (Delay\_Total  $\rightarrow$  Delay\_Rate) etc., are disclosed towards the ignorance in structure (a) of Figure 4. And this also happens to our first experiment as shown in Figure 4. It demonstrates that PNM makes the resulting BN more sensitively recognize the correlation in the data.

## 5. CONCLUSIONS AND FUTURES WORK

In this paper, we have analyzed most previous methods of data approximation, and point out that the following two basic principles shouldn't be ignored, as we approximate the data for BN. Both frequency and value optimization are necessary for data approximation. For Bayesian network, principle 1 is working in probability space. From these principles, we have defined a norm minimum criterion Eq.(6) in probability space (PNM), and solve the equation by a kernel based function as Eq.(9), and realize this minimizing procedure by a iterative approximating algorithm. Our accuracy and convergency tests on this algorithm show that the proposed PNM approximation is practically converged and its simulation accuracy is superior to the popularly used histogram method.

We have implemented PNM criterion to BN learning on both the UCI database and real database. The performance of the trained Naïve-Bayes classifier demonstrates that our method is more effectively to fit the data than the work of (EL-Matouat, F. et. al. 2000). Furthermore, by the contrast results of BN structure modeling, it has been proved that a PNM optimal data approximation is advantageous for BN to recognize the correlation in the data, and to make the resulted BN fit data and prior knowledge better.

In this work, the resolution of data approximation plays an important role in BN learning. The resolution adjustment of approximation makes a new BN topic of multi-resolution analysis data. We expect it to be a valuable topic, and the source of much future work.

## REFERENCES

- Campos De etc. al. (2000), Building Bayesian network-based information retrieval systems, Proc. 11th International Workshop on Database and Expert Systems Applications, 543 -550.
- Geiger D. (1992), An Entropy-Based Learning Algorithm of Bayesian conditional Trees," Duubois et al. (eds.), Proc. Eighth conf on Uncertainty in Artificial Intelligence, Standford, Calif..
- Kullback S. (eds.)(1959), *Information Theory and Statistics*, John Wiley & Sons, New York.
- Poland R.B. and Shachter R.D. (1994), Three Approaches to Probability Model Selection, R.Lopez de Mantaras and D. Poole (eds.), Proc. Ninth Conf. on Uncertainty in Artificial Intelligence, Seattle Wash., 478-483.

- Sclove S.L. (1994), Small-Sample and Large-Sample Statistical Model Selection Criteria, P. Cheeseman and R.W. Oldford (eds.), *Selecting Models from Data: Artificial Intelligence and Statistic IV*, Springer-Verlog, 31-39.
- EL-Matouat, F. et. al. (2000), From continuous to Discrete Variables for Bayesian Network Classifiers, Proc. IEEE Inter. Conf. on Systems, Man, and Cybernetics, Vol.4 , 2800 –2805.
- Cooper G. F. and Herskovitz E. (1992) “A Bayesian method fro the induction of probabilistic networks from data,” *Machine Learning* , 9 , 309-347.
- Cowell R.G., Dawid A.P., Laurithzen S.L., and Spiegelhalter D.J. (1999), *Probabilistic Networks and Expert Systems*, Springer-Verlag, New York.
- Herskovits E.H. and Cooper G.F. (1995), Kutató: An Entropy-Driven System for Construction of Probabilistic Expert systems From Databases, Piero Bonissone, (eds.), Proc. Sixth Conf. Uncertainty in Artificial Intelligence, Montreal, 54-62.
- Geiger D. (1992), An Entropy-Based Learning Algorithm of Bayesian Conditional Trees, Dubois et al. (eds.) Proc. Eighth Conf. on Uncertainty in Artificial Intelligence: Stanford, Calif., 92-97.
- Viswanath P., Yannis E. Ioannidis, Peter J. Haas and Eugene J. Shekita, (1996), Improved Histograms for Selectivity Estimation of Range Predicates, SIGMOD’96, 294-305.
- Ioannidis Y. and Poosala V. (1995), Balancing histogram optimality and practicality for query result size estimation., Proc. of ACM SIGMOD Conf, 233-244.
- Haas P.J. and Swami A.N. (1995), Sampling-based selectivity estimation for joins using augmented frequent value statistics, Prod. of IEEE Conf. On Data Engineering, 522-531.
- Korn, F., Johnson, T. and Jagadish, H.V. (1999), Range selectivity estimation for continuous attributes”, Proc. Of Eleventh Inter. Conf. on Scientific and Statistical Database Management, 244 –253.
- Wray Buntine, (1996), “A Guide to the Literature on Learning Probabilistic Networks from Data,” *IEEE Transactions on Knowledge and Data Engineering*, 8(2), 195-210.
- Kirsch, H. and Kroschel, K. (1994), Applying Bayesian networks to fault diagnosis, Proceedings of the Third IEEE Conference on Control Applications, Vol. 2, 895 -900.
- Suojanen, M., Andreassen, S. and Olesen, K.G. (2001), “A method for diagnosing multiple diseases in MUNIN”, *IEEE Transactions on Biomedical Engineering*, **48** (5) , 522 –532.

**Received:** May 10<sup>th</sup> 2004

**Accepted in final format:** Nov 20<sup>th</sup> 2004

**About the author:**

Dr. Shaoning Pang received his B. S. in physics, M. S. in electronic engineering and Ph.D. in computer science. From 2001 to 2003, he worked as a research associate in Pohang University of Science and Technology (POSTECH), South Korea. Dr. Pang is a member of IEEE, IEICE, and ACM. He also served as a reviewer for IEEE trans. on SMC-Part B, and Pattern Recognition Letter. His research interests include Support Vector Machines, Incremental Learning, and Bioinformatics. He can be reached by email at [spang@aut.ac.nz](mailto:spang@aut.ac.nz)