International Journal of Fuzzy Logic and Intelligent Systems, vol. 10, no. 4, December 2010, pp. 253−258
DOI : 10.5391/IJFIS.2010.10.4.253

# An Extended Version of the CPT-based Estimation for Missing Values in Nominal Attributes

**Song Ko and Daewon Kim**

**School of Computer Science and Engineering, Chung-Ang University, 221 Heukseok-dong, Dongjak-gu, Seoul 156-756, Korea**

## Abstract

The causal network represents the knowledge related to the dependency relationship between all attributes. If the causal network is available, the dependency relationship can be employed to estimate the missing values for improving the estimation performance. However, the previous method had a limitation in that it did not consider the bidirectional characteristic of the causal network. The proposed method considers the bidirectional characteristic by applying prior and posterior conditions, so that it outperforms the previous method.

**Key Words** :Causal Network, Missing Values, Dependency Relationship, Estimation

## 1. Introduction

Most nominal datasets encountered in practice contain missing values [1, 2]. Attribute values unobserved from various causes are often indicated by blanks or '?'. Many methods to estimate missing values have been researched [2, 3, 4, 5]: among them, probability-based estimations are the most widely used.

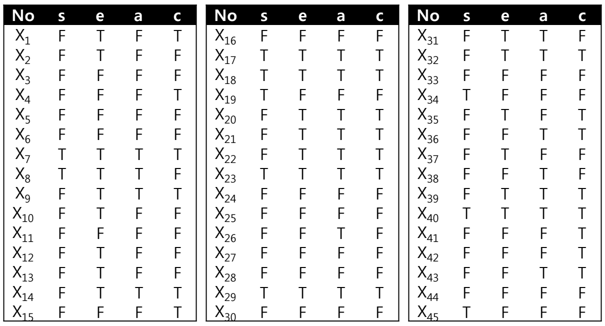
We use an example data to explain the probability-based method. Figure 1 contains 45 patterns with four nominal attributes; **s**(steal), **e**(earthquake), **a**(alarm) and **c**(call). $X_i$ = [$X_i$(**s**), $X_i$(**e**), $X_i$(**a**), $X_i$(**c**)]$^T$ represent $i$-th pattern ($1 \leq i \leq 45$), and each nominal attribute has a value of (T)rue or (F)alse; i.e., $X_1(\mathbf{a} = \mathtt{T})$ denotes that $X_1$'s value for 'alarm' attribute is T. Let us suppose that five values for the attribute **a** were missing; in Fig. 2(a), $X_7$,$X_{22}$,$X_{30}$,$X_{37}$,$X_{41}$ have missing values symbolized with '?'. The notation of $X_7$(**a**=?) denotes that the value **a** for $X_7$ is missing. To deal with $X_7$(**a**=?), the probability-based estimation calculates a marginal probability table (MPT). Given number of patterns ($n$) and number of the missing patterns ($\beta$), the probability of $X_i$(**a**=T), denoted by P($X_i$(**a**=T)), is calculated by:

$$\begin{aligned} P(X_i(\mathbf{a} = \mathtt{T})) &= \frac{1}{n-\beta}\sum_{k=1}^{n} I(X_k(\mathbf{a} = \mathtt{T})) \\ &= \frac{1}{45-5}\sum_{k=1}^{45} I(X_k(\mathbf{a} = \mathtt{T})) = \frac{17}{40} = 0.425 \end{aligned} \tag{1}$$

where $I(\cdot)$ function returns 1.0 when $X_k$'s value for 'alarm' is T; otherwise; it returns 0.0. From Eq. 2, we see that 17 patterns have $X_i$(**a**=T), and P($X_i$(**a**=T)) equals to 0.425; similarly, P($X_i$(**a**=F)) equals to 0.575 (=23/40). We can compute the marginal probability table for other attributes (**s**, **e** and **c**) similarly, shown in Fig. 2(b). A MPT-based estimation imputes the missing values based on the marginal probability table(Fig 2.(b)). Thus, in the MPT-based estimation, $X_7$(**a**=?) is finally estimated to be F, since $X_7$(**a**=F) has a higher probability than $X_7$(**a**=T); P($X_7$(**a**=F))=0.575 $>$ P($X_7$(**a**=T))=0.425. Similarly, all other missing values of **a** are estimated to be F. We find that two of the five missing patterns were incorrectly estimated, since $X_7$ and $X_{22}$ have a value of T in the original data (Fig. 1).

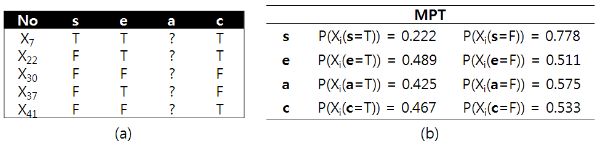
Marco et al. used the dependency relationship between attributes to improve estimation performance, when a causal network is available for the attributes [6]. In the causal network of Fig. 3(a), we see that **a** is influenced by its two parents (**e** and **s**). By employing the parents (or prior conditions), Marco et al. improved the estimation performance. This method is termed CPT (conditional probability table) based estimation (or prior condition based estimation). To deal with $X_{22}$(**a**=?), we need to construct the CPT, considering **s** and **e**, which are parents of **a**, to estimate the missing values of **a**. For example, the probability of $X_i$(**a**=T) under the prior conditions (**s**=F and **e**=T) is calculated by:

Manuscript received Jul. 28, 2010; revised Dec. 2, 2010;
Accepted Dec. 10, 2010
Corresponding Author : dwkim@cau.ac.kr (Daewon Kim)



| No | s | e | a | c | No | s | e | a | c | No | s | e | a | c |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| $X_1$ | F | T | F | T | $X_{16}$ | F | F | F | F | $X_{31}$ | F | T | T | F |
| $X_2$ | F | T | F | F | $X_{17}$ | T | T | T | T | $X_{32}$ | F | T | T | T |
| $X_3$ | F | F | F | F | $X_{18}$ | T | T | T | T | $X_{33}$ | F | F | F | F |
| $X_4$ | F | F | F | T | $X_{19}$ | T | F | F | F | $X_{34}$ | T | F | F | F |
| $X_5$ | F | F | F | F | $X_{20}$ | F | T | T | T | $X_{35}$ | F | T | F | T |
| $X_6$ | F | F | F | F | $X_{21}$ | F | T | T | T | $X_{36}$ | F | F | T | T |
| $X_7$ | T | T | T | T | $X_{22}$ | F | T | T | T | $X_{37}$ | F | T | F | F |
| $X_8$ | T | T | T | F | $X_{23}$ | T | T | T | T | $X_{38}$ | F | F | T | F |
| $X_9$ | F | T | T | T | $X_{24}$ | F | F | F | F | $X_{39}$ | F | T | T | T |
| $X_{10}$ | F | T | F | F | $X_{25}$ | F | F | F | F | $X_{40}$ | T | T | T | T |
| $X_{11}$ | F | F | F | F | $X_{26}$ | F | F | T | F | $X_{41}$ | F | F | F | T |
| $X_{12}$ | F | T | F | F | $X_{27}$ | F | F | F | F | $X_{42}$ | F | F | F | T |
| $X_{13}$ | F | T | F | F | $X_{28}$ | F | F | F | F | $X_{43}$ | F | F | T | T |
| $X_{14}$ | F | T | T | T | $X_{29}$ | T | T | T | T | $X_{44}$ | F | F | F | F |
| $X_{15}$ | F | F | F | T | $X_{30}$ | F | F | F | F | $X_{45}$ | T | F | F | F |

Fig. 1 An example original dataset of 45 patterns with four nominal attributes.



| No | s | e | a | c |
|---|---|---|---|---|
| $X_7$ | T | T | ? | T |
| $X_{22}$ | F | T | ? | T |
| $X_{30}$ | F | F | ? | F |
| $X_{37}$ | F | T | ? | F |
| $X_{41}$ | F | F | ? | T |

(a)

| | MPT | |
|---|---|---|
| **s** | P(X$_i$(**s**=T)) = 0.222 | P(X$_i$(**s**=F)) = 0.778 |
| **e** | P(X$_i$(**e**=T)) = 0.489 | P(X$_i$(**e**=F)) = 0.511 |
| **a** | P(X$_i$(**a**=T)) = 0.425 | P(X$_i$(**a**=F)) = 0.575 |
| **c** | P(X$_i$(**c**=T)) = 0.467 | P(X$_i$(**c**=F)) = 0.533 |

(b)

Fig. 2 (a) Five patterns have missing values for the attribute 'alarm'; (b) a marginal probability table of the four attributes

$$
\begin{aligned}
&P(X_i(\mathbf{a}=\mathtt{T})|X_i(\mathbf{s}=\mathtt{F}),X_i(\mathbf{e}=\mathtt{T}))\\
&\quad=\frac{1}{m}\sum_{k=1}^{n}I(X_k(\mathbf{a}=T)|X_k(\mathbf{s}=\mathtt{F}),X_k(\mathbf{e}=\mathtt{T}))\\
&\quad=\frac{8}{15}=0.533 \qquad (2)
\end{aligned}
$$

where $m$(=15) is the number of patterns corresponding to the given prior conditions (**s**=F and **e**=T). $I(\cdot)$ function returns 1.0 when X$_k$'s value for 'alarm' is T under the two conditions; otherwise; it returns 0.0. We see that eight patterns have X$_i$(**a**=T), and P(X$_i$(**a**=T)|X$_i$(**s**=F), X$_i$(**e**=T)) equals to 0.533; similarly P(X$_i$(**a**=F)|X$_i$(**s**=F), X$_i$(**e**=T)) equals to 0.467. Fig. 3(b) represents the CPT of **a**. Thus, X$_{22}$(**a**=?) is estimated by T, since P(X$_{22}$(**a**=T)) has a higher probability than P(X$_{22}$(**a**=F)) under the prior conditions; P(X$_{22}$(**a**=T))=0.533 $>$ P(X$_{22}$(**a**=F))=0.467. We find that one of the five missing patterns, $\mathbf{X}_{37}$, is still incorrectly estimated.

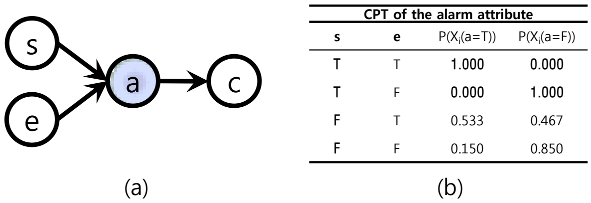
## 2. Proposed Methods

In this study, we extend the CPT-based estimation using the bidirectional characteristic of causal networks. The bidirectional characteristic means that a event is influenced by prior conditions and influences to posterior conditions. Fig. 4 shows an illustration for this concept, which represents the causal network for dependencies between a recent visit to Korea and the chances of dyspnoea (shortness-of-breath). Let us suppose that some missing values were occurred in *Lung cancer* and we try to estimate them. Under this case, *Lung cancer* is influenced by the prior condition (*Smoking*) and influences to the posterior condition (*Abnormality in Chest*). Thus, the estimation for missing values in *Lung cancer* can be improved by exploiting both prior and posterior conditions together.

Similarly, we can see from Fig. 5 that 'alarm'(**a**) is influenced by prior conditions ('steal' and 'earthquake') and influences to the posterior condition ('call'). Thus, it is possible to improve the estimation of missing values of **a** by considering the two-types of conditions. However, to our knowledge, few research considers the posterior condition being used to estimate the missing values. Therefore, we proposed a method to consider the posterior condition to estimate the missing values with the prior conditions simultaneously. The proposed method is termed E-CPT(Extended CPT) based estimation.

First, we examine the case that considers only the posterior condition to estimate the missing values (Fig. 5(b); Let us assume that the **a** has one child with no parent). To estimate the missing values of **a**, we need to construct the probability table of **a** considering the posterior condition **c**. The probability of X$_i$(**a**=T) under the posterior condition(**c**=T) is calculated by:

$$
\begin{aligned}
&P(X_i(\mathbf{a}=\mathtt{T})|X_i(\mathbf{c}=\mathtt{T}))\\
&\quad=\frac{1}{m}\sum_{k=1}^{n}I(X_k(\mathbf{a}=\mathtt{T})|X_k(\mathbf{c}=\mathtt{T}))\\
&\quad=\frac{15}{21}=0.714 \qquad (3)
\end{aligned}
$$



| CPT of the alarm attribute | | | |
|---|---|---|---|
| s | e | P(X$_i$(a=T)) | P(X$_i$(a=F)) |
| T | T | 1.000 | 0.000 |
| T | F | 0.000 | 1.000 |
| F | T | 0.533 | 0.467 |
| F | F | 0.150 | 0.850 |

Fig. 3 (a) a causal network of dependency relationship between attributes; (b) a conditional probability table of the alarm attribute

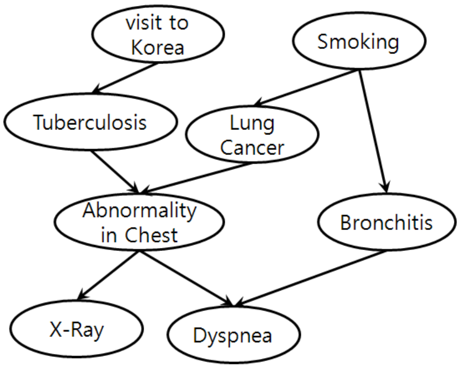


Fig. 4 Causal Network for the Asia data set

where $m$(=22) is the number of patterns corresponding to the given posterior condition. 15 patterns have X$_i$(**a**=T) under **c**=T, and P(X$_i$(**a**=T)|X$_i$(**c**=T)) equals to 0.714; similarly P(X$_i$(**a**=F)|X$_i$(**c**=T)) equals to 0.286. From Fig. 6(a), X$_7$(**a**=?) is estimated as T, since X$_7$(**a**=T) has a higher probability than X$_7$(**a**=F) under the posterior condition; P(X$_7$(**a**=T))=0.714 $>$ P(X$_7$(**a**=F))=0.286. X$_{30}$(**a**=?) is estimated as F under **c**=F; P(X$_{30}$(**a**=T))=0.167 $<$ P(X$_{30}$(**a**=F))=0.833. We find that one of the five missing patterns is incorrectly estimated; $X_{41}$ is estimated as T, although it is F in the original data. We can find that the estimation result that considers the posterior conditions performs similarly to that of the typical CPT based estimation.

The proposed E-CPT based method improves the estimation performance of the missing values by considering both prior and posterior conditions. In the example data, **a** has three conditions that consist of the two prior conditions(parents; **s** and **e**) and one posterior condition(child; **c**). We need to construct the E-CPT of **a** considering all conditions(**s**, **e** and **c**) to estimate the missing values at the **a**. For example, P(X$_i$(**a**=T)) under the conditions (**s**=F, **e**=T, and **c**=T) is calculated by:

$$
\begin{aligned}
&P(X_i(\mathbf{a}=\mathtt{T})|X_i(\mathbf{s}=\mathtt{F}), X_i(\mathbf{e}=\mathtt{T}), X_i(\mathbf{c}=\mathtt{T})) \\
&\quad = \frac{1}{m}\sum_{k=1}^{n} I(X_k(\mathbf{a}=\mathtt{T})|X_k(\mathbf{s}=\mathtt{F}), X_k(\mathbf{e}=\mathtt{T}), X_k(\mathbf{c}=\mathtt{T})) \\
&\quad = \frac{7}{9} = 0.778 \qquad (4)
\end{aligned}
$$

Nine ($m$=9) patterns correspond to the given conditions, seven of which have X$_i$(**a**=T). Therefore, P(X$_i$(**a**=T)|X$_i$(**s**=F), X$_i$(**e**=T), X$_i$(**c**=T)) equals to 0.778. Fig. 6(b) shows the E-CPT of **a**. Therefore, X$_{22}$(**a**=?) is estimated as T, under the three conditions; P(X$_{22}$(**a**=T))=0.778 $>$ P(X$_{22}$(**a**=F))=0.222. X$_7$(**a**=?) is estimated as T, since X$_7$(**a**=T) has a higher probability than X$_7$(**a**=F), under the conditions (**s**=T, **e**=T, and **c**=T); P(X$_7$(**a**=T))=1.000 $>$ P(X$_7$(**a**=F))=0.000. Finally, we can state that no incorrect estimation result appears in the example data.

## 3. Results

### 3.1 Experimental Results

To show the effectiveness of the proposed method, we compared the E-CPT based method with the MPT and CPT based imputation methods and KNN imputation method

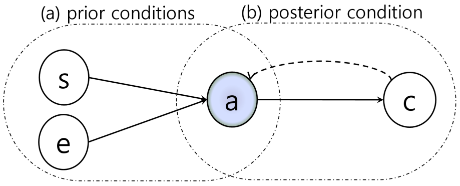
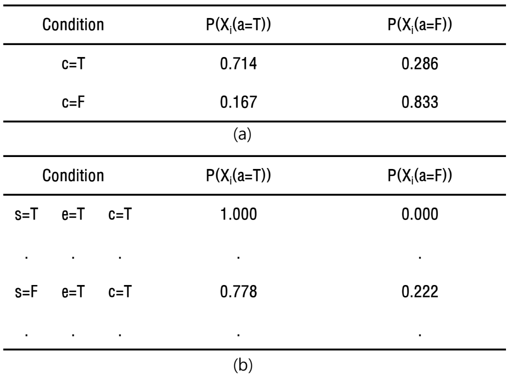


Fig. 5 The bidirectional characteristic of the causal Network. The solid lines represents prior condition based estimation (or CPT based estimation); the dotted line represents posterior condition based estimation.



| Condition | $P(X_i(a{=}T))$ | $P(X_i(a{=}F))$ |
|---|---|---|
| c=T | 0.714 | 0.286 |
| c=F | 0.167 | 0.833 |

(a)

| Condition | | | $P(X_i(a{=}T))$ | $P(X_i(a{=}F))$ |
|---|---|---|---|---|
| s=T | e=T | c=T | 1.000 | 0.000 |
| . | . | . | . | . |
| s=F | e=T | c=T | 0.778 | 0.222 |
| . | . | . | . | . |

(b)

Fig. 6 (a) Probabilities of $X_i(\mathbf{a=T})$ and $X_i(\mathbf{a=F})$ using a posterior condition; (b) the extended version of CPT for the alarm attribute under three conditions

where K = 3, 5. We experimented with two data sets; Car Start data and Asia data. These data are available on the BayesiaLab web site at http://www.bayesia.com. The Car Start data set has 10,000 patterns with 18 attributes. The Asia data set has 10,000 patterns with eight attributes (Fig. 4). We randomly create from 5% to 35%, at 5% increments, missing values. We evaluate the average estimation performance after 30 repetitions at each missing rate. The estimation accuracy were assessed using the ratio of the number of correctly estimated data to the total number of missing data. A higher value of accuracy shows a better estimation.

Fig. 7 represents the performance of the three methods for the Asia data set. At the 15% missing rate, MPT based estimation has 76.2% performance compared to CPT's 81%. The proposed method exhibits the best performance by 84.5%. Overall, the MPT based estimation exhibits the worst performance of 76% among the probability based imputation, regardless of the missing rate. The CPT based estimation exhibits about 79% $\sim$ 81% and improves by about 3% $\sim$ 5% compared to MPT based estimation. The proposed method exhibits the best performance with 82% $\sim$ 85% and improves about 6% $\sim$ 9% compared to MPT based estimation. On the other hand, the KNN imputation method shows similar accuracy with the proposed method at the lower missing rates, especially at the 5% missing rate. However, as the missing rate is increased, the performance is significantly decreased. Fig. 8 represents the performances of the all methods for the Car Start data. At the 15% missing rate, MPT based estimation exhibits 89.5% performance, CPT based estimation exhibits 93% of performance and the proposed method has the best performance, of 95.5%. Overall, MPT based estimation exhibits the worst performance at 89%. CPT based estimation exhibits about 92% and improves by about 3% performance compared to MPT based estimation. The KNN methods inducts similar results at the Asia data. Although the KNN inducts better performance at the 5% missing rate, the performance is significantly decreased as increasing the missing rate. The proposed method improves performance by 5% $\sim$ 7%; it exhibits the best performance of about 95%.

We analyze the estimation performance of the two specific attributes; *Bronchitis*, which has one child and one parent, and *Smoking*, which has only one child with no parent(Fig. 9). For *Bronchitis*, MPT based estimation exhibits the worst performance of about 56.4%. CPT based estimation improves estimation by considering *Smoking*, which is the parent of *Bronchitis*, with 64.9%. Notably, the best estimation performance is induced by the proposed method with 85.1% performance, since two conditions, *Smoking* and *Dyspnea* which is the child of the *Bronchitis*, are employed to estimate the missing values. For *Smoking*, MPT based estimation exhibits about 53.5% performance and CPT based estimation exhibits about 52.9% performance. There is no difference in the estimation performances of the two estimations, since *Smoking* has no parent. However, by

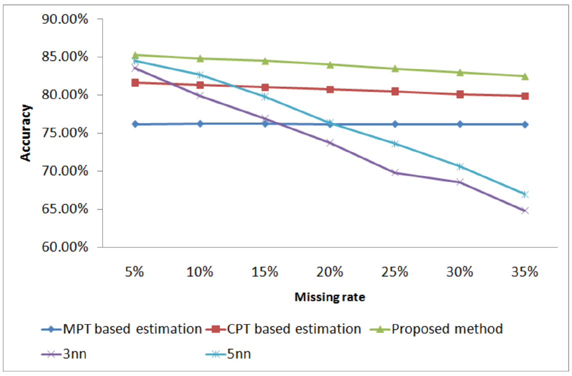


Fig. 7 The performance comparison of the three estimation methods for the Asia data

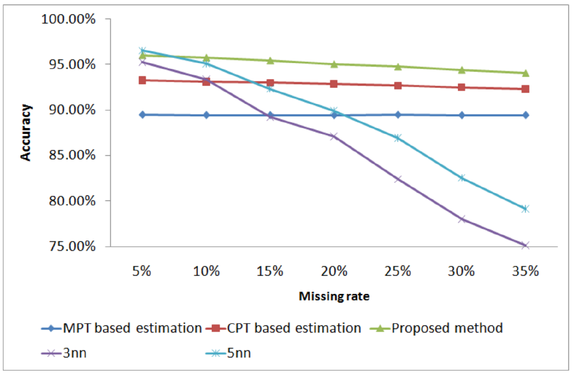


Fig. 8 The performance comparison of the three estimation methods for the Car Start data

considering the child of the *Smoking*, the performance of the proposed method is improved to 65.8%.

## 3.2 Analysis

We have showed that employing both the prior and posterior conditions is helpful to improve the estimation performance despite the causal relationship was broken. It supports that the causal relationship can be employed alternatively depending on the problem domain. The estimation problem of the missing value does not need to employ the causal relationship, in this case using only prior condition, but employ the both conditions[7]. There are several reasons at the improvement of the estimation performance when employing the both conditions: The first, the attributes having the causal relationship are not independent each other. Therefore, they have high value of the joint probability distribution and it helps to improve the estimation performance. The second, the likelihood equivalent property can be adjustable for estimating the missing value of the specific attributes. Estimation method employs the part of the causal network directly related to the specific attributes. In this case, the attributes are helpful to improve the estimation performance if they have the causal relationship whether they are parents or children[7].

## 4. Discussion

In this paper, we demonstrated that E-CPT based missing values estimation outperforms previous methods, such as CPT based estimation, since E-CPT based estimation considers the bidirectional characteristic of the causal Network. Finally, we can see that both the prior and the posterior conditions are available to improve the estimation performance for the missing values.

**Acknowledgement.** This research was supported by Basic Science Research Program through the National Research Foundation of Korea(NRF) funded by the Ministry of Education, Science and Technology(2010-0012885).

## References

[1] A.Rogier T.Donders, Geert J.M.G. van der Heijden, Theo Stijnen, Karel G.M. Moons, Review: “A gentle introduction to imputation of missing values”, *Journal of Clinical Epidemiology* Vol.59, pp.1087-1091, 2006.

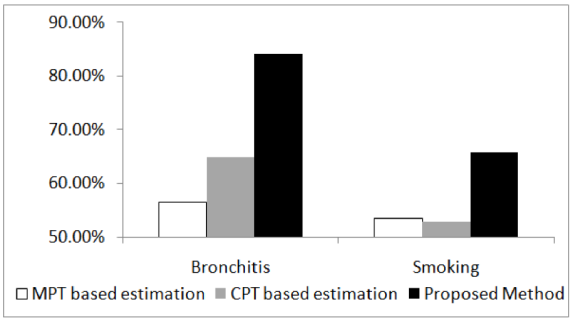


Fig. 9 Experimental results when missing values occur for the *Bronchitis* attribute and the *Smoking* attribute in the Asia data set.

[2] Alireza Farhangfar, Lukasz A.Kurgan, Member, IEEE, and Witold Pedrycz, Fellow, IEEE, “A Novel Framework for Imputation of Missing Valuesin Databases”, *IEEE Transaction on Systems*, Man, and Cybernetics-PART A : System and Humans, Vol.37, NO.5, pp.692-709, SEPTEMBER 2007.

[3] Marco Ramoni and Paola Sebastiani, “Robust Learning with Missing Data”, *Machine Learning*, Vol.45, pp.147-170, 2001.

[4] S.F.Buck, “A Method of Estimation of Missing Val-ues in Multivariate Data suitable for use with an Electronic Computer”, *Journal of the Royal Statisti-cal Society. Series B* (Methodological), Vol.22, NO.2, pp.302-306, 1960.

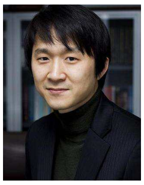
[5] Z.Ghahramani and M.I.Jordan, “Mixture models for learning from incomplete data”, *Computational learning theory and natural learning systems* : VolumeIV, MIT Press, pp.67-85,1997.

[6] Marco Di Zio, Mauro Scanu, Lucia Coppola, Orietta Luza and Alessandra Ponti, “Bayesian Network for Imputation”, *Journal of the Royal Statistical Society* : SeriesA, Vol.167, Part2, pp.309-322, 2004.

[7] Shulin Yang and Kuo-Chu Chang, “Comparison of Score Metrics for Bayesian Network Learning”, *IEEE Transaction on Systems*, Man and Cybernetics-PARTA : Systems and Humans, Vol.32, No.3, pp.419-428, May2002

[8] Judea Pearl, *Probabilistic Reasoning in Intelligent Systems : Network of Plausible Inference*, 1988.

**Song Ko** received the M.S. degree in Computer Science and Engineering, Chung-Ang University, Korea. He is currently in the Ph.D course. His research interests include data mining, Bayesian networks, knowledge representation.



**Daewon Kim** received the M.S. and Ph.D degrees in computer science from Korea Advanced Institute of Science and Technology(KAIST), Daejon, Korea. in 1999 and 2004. Since 2005, he has been an assistant professor at the School of Computer Science and Engineering, Chung-Ang University, Korea. His research interests include data mining, pattern recognition, and artificial intelligence.