Kuwait J. Sci. **40** (2) pp. 83-101, 2013

Improving the performance of Bayesian networks in non-ignorable missing data imputation

P. NILOOFAR, M. GANJALI AND M.R. FARID ROHANI

Department of Statistics, Faculty of Mathematical Science, Shahid Beheshti University, Tehran, Iran. Email: Pniloofar@yahoo.com, m-ganjali@sbu.ac.ir, M_faridrohani@sbu.ac.ir

ABSTRACT

The issue of missing data may arise for researchers who deal with data gathering problems. Bayesian networks are one of the proposed methods that have been recently used in missing data imputation. The main objective of this research is to improve the efficiency of the Bayesian networks in nonignorable missing imputation, by adding missing indicator nodes for incomplete variables and constructing an augmented Bayesian network. Also, to consider the effect of different kinds of missingness mechanism (ignorable and nonignorable) on the performance of imputation methods. Four methods of imputation: random overall hot-deck imputation, within-class random hot-deck imputation, imputation using Bayesian networks and imputation using presented augmented Bayesian networks are compared using two indices: (1) a distance function and (2)Minimum Kullback-Leibler index. Results indicate the high-quality of the methods based on Bayesian networks relative to other imputation methods.

Keywords: Bayesian networks; imputation; kullback-Leibler information; nonignorable mechanism; value of information analysis.

INTRODUCTION

The issue of missing data may arise for researchers who deal with data gathering problems. According to (Rubin, 1976), the assumptions about the missing data mechanisms may be classified into three categories: (1) missing completely at random (MCAR):the probability that an entry is missing is independent of both observed and unobserved values in the data set. For example a patient may simply forget to post the questionnaire back; (2) missing at random (MAR): the probability that an entry is missing is a function of the observed values in the data set; (3) informatively missing (IM) or Non-MAR (NMAR): the probability that an entry is missing depends on both observed and unobserved values in the data set. An example is that a person with reduced health condition due to side effects of a treatment may be less likely to return the questionnaire. In this case, simply excluding those with missing data from the analysis, will bias the results if those who did not respond were in significantly lower (or higher) health condition than those who did respond.When data are either MCAR or MAR, the deletion mechanism is said to be ignorable because we can infer the missing entries from the observed ones. A problem in coping with missing imputation is the preservation of joint relationships between variables.

Di Zio *et. al.* (2004) developed an imputation method based on the revised version of the method proposed by Thibaudeau & Winkler (2002). Its main goal was to use Bayesian networks for imputation tasks in order to preserve joint distributions as much as possible. Unfortunately, their method is based on the assumption that the mechanism of missing data is not IM. This assumption is hard to test in practice and the decrease in accuracy may be severe when the assumption is violated.

Recently, Lin & Haug (2008) experimented a method of treating missing values in a clinical data set by explicitly modeling the absence of data. They showed that in most cases, a naive Bayesian network trained using the explicit missing value treatments performed better. Also some approaches have been designed with a view to be ‘robust’ to the missing data mechanism (Ramoni & Sebastiani, 2001; Aussem & Rodrigues de Morais, 2010). In Ramoni & Sebastiani (2001) method, called Robust Bayesian Estimator (RBE), no assumption about the unknown censoring mechanism is made. Rodrigues de Morais & Aussem (2009) exploited the missing mechanism using both the Bayesian networks and the information represented by the absence of data and they reduced the classification error.

In this study, the main objective is to improve the efficiency of the Bayesian networks in the task of imputation, especially when the missingness pattern is IM. We tried to reach this goal, by introducing a novel approach based on an augmented Bayesian network. To assess the benefits and drawbacks of this approach, an experimental study was conducted on two data sets. First data set is extracted through an experimental study on a data set of individual records obtained from the Iran Statistical Research Centre, and the second data set is extracted from the study that examines a sample of 405 children within the first two years of entry to elementary school. We also compare the performance of the methods based on Bayesian networks (the method introduced by Di Zio *et. al.*, 2004, and our augmented Bayesian network) with that of random overall hot-deck imputation and that of within-class random hot-deck imputation with respect to different stratifications, by a simulation study. In this paper, to evaluate and compare the performance of these applied methods, in addition to the delta index of Di Zio *et. al.* (2004, 2005), we will also introduce the use of MinimumKullback-Leibler index. A simulation study and a value of information analysis will be performed to investigate the robustness and sensitivity of the Bayesian network with respect to the different mechanisms of missing data.

In Section 2, a brief overview of Bayesian networks is presented. Section 3, is devoted to the description of the use of Bayesian networks for imputation, introducing the augmented Bayesian network approach, and the definition of the measures which will be used for evaluation. Information evaluation and explanations of the applied experiments are given in Sections 4 and 5, respectively. Section 6 includes the results obtained and conclusions.

BAYESIAN NETWORKS

A Bayesian network is a graphical model that encodes probabilistic relationships among variables of interest and is defined by three elements (Jensen, 1996):

- The nodes: each node represents a variable with a finite number of states
- The directed edges: Each edge connects a pair of nodes
- To each variable is attached a conditional probability distribution.

The first two points define the qualitative (structure) part of the Bayesian network and the third defines the quantitative (parameter) thereof. The definition of Bayesian networks does not refer to causality, and there is no requirement that the links represent causal impact. That is when building the structure of a Bayesian network model, it is not required to insist on having the links go in a causal direction (Jensen & Nielsen, 2007). However, the model should not include conditional independences that do not hold in the real world.

By the chain rule given in Pearl (1988), the following formula for the complete joint probability distribution for a Bayesian network can be derived:

$$P(X_1 = x_1, X_2 = x_2, ..., X_k = x_k) = \prod_{i=1}^{k} P\{X_i = x_i | Pa(X_i)\} \qquad (1)$$

Where $Pa(X_i)$ is the set of variables immediately preceding the variable X_i. This set is called the parent of the variable X_i. So, to each variable is attached a probability distribution conditioning on its parents. This formulation leads to a dramatic decrease in the number of parameters to be estimated (Lauritzen & Spiegelhalter 1988; Cowell *et. al.*, 1999).

METHOD

This section outlines the method needed to use Bayesian networks for imputation. First, a Bayesian network must be specified for the incomplete data set, and then this specified Bayesian network will be used to impute missing items. At the end of this section, the performance of the applied Bayesian network will be evaluated.

Learning Bayesian Networks from data

The most prominent feature of Bayesian networks is their ability to learn from data, either parameters (i.e. conditional probabilities) or learning structure or both. Indeed, considering whether a data set is complete or not, and whether the network structure is known or not, there are four possibilities for a Bayesian network to be learnt from data as shown in Table 1. This study falls into the cell categorized by ‘Incomplete data’ and ‘Unknown structure’, other possibilities refer to Buntine (1994) and Heckerman (1996).

Table 1. Four possibilities for a Bayesian network to be learnt from data (the one we deal with indicated by √)

Bayesian Network Structure	**Data**	
	Complete	**Incomplete**
Known		
Unknown		√

The unknown structure part is the most usual one (Cowell *et. al.*, 1999). Furthermore, estimating a Bayesian network from a data set comprises two steps:

- Structure learning
- Parameter learning.

In this paper, PC algorithm (Spirtes *et. al.*, 1993) is used for the structure learning step (Algorithm 1).

```
1. Start with the complete graph;
2. $i := 0$;
3. While a node has at least $i+$ 1neighbors
       - For all nodes $A$ with at least $i + 1$ neighbors
       - For all neighbors $B$ of $A$
       - For all neighbor sets $\chi$ such that $|\chi| = i$ and $\chi \subseteq (neighbor(A)\backslash\{B\})$
       - If $A$ and $B$ are conditionally independent given $\chi$ then remove link $A - B$
       - $i := i + 1$
```

Algorithm 1: The PC algorithm: test sequence

After the structure learning step, we may condition on it and estimate the joint distribution of the variables by maximum likelihood estimation. When missing items are present, the EM algorithm is used. In this research, the procedure described by Lauritzen (1995) is applied to the parameter learning step.

Imputation based on Bayesian networks

Consider a data set with K variables of interest, $X_1, X_2, ..., X_k$. Among these, some variables may be completely observed (O), and some variables may have missing items (M). We denote the whole data set by D = (O,M).The structure of a Bayesian network defines a hierarchical ordering among the variables $X_1, X_2, ...X_k$. As was stated in Section 2, the structure of a Bayesian network is defined by the edges and their direction. Although edges' direction is usually interpreted in terms of causal relationships between two nodes, Di Zio *et.al.* (2004) apply a different interpretation: at first, the variables are ordered according to their 'reliability'. More reliable variables are those with a lower percentage of missing items, higher accuracy and greater availability of external sources. Once a direct statistical relationship between a pair of variables has been found, the edges' direction is defined from the most reliable variable to the least. This ordering defines a partition of the data set D = (O,M) in v mutually exclusive subsets $P_j, j = 0, ..., v - 1$ where v is the number of variables in the longest chain. The first subset P_0 contains variables with no parents. The second subset P_1 includes the remaining variables whose parents are only in P_0. The third subset P_2 contains the remaining variables whose parents are only in $P_0 \cup P_1$. Generally, once the first $j - 1$ subsets have been established, the jth subset P_j contains the remaining variables whose parents are only in $\bigcup_{h=0}^{j-1} P_h,\ j = 1, ..., v$ (Di Zio *et. al.*, 2004).

According to Thibaudeau & Winkler (2002), the previous ordering and the distribution of the corresponding Bayesian network are crucial for the following imputation procedure. Assume a sample of n units, $a = 1, ..., n$, the variables $X_1, X_2, ..., X_k$ are collected and there are missing items. For the first unit, $a = 1$, we check the presence of a missing value starting from the variables in P_0. If a variable X in P_0 is missing, X is imputed by randomly generating a value from the marginal distribution of X. If the variables in $P_{j-1}, (j > 1)$ are either present or imputed and a variable X in P_j is missing, X is imputed by randomly generating a value from the distribution of X conditionally on the values assumed by its parents. Such a procedure is iterated for all the variables in the v groups and all the units $a = 1, ..., n$ (Di Zio *et.al.*, 2004; Di Zio *et. al.*, 2005).

Imputation based on Augmented Bayesian networks

According to the method described in section 3.2, first a Bayesian network must be specified for the original data set D = (O,M), and then the constructed Bayesian network will be used to impute the missing items (BN method). But this method relies on the assumption that the missing pattern should be ignorable (MCAR or MAR).

In order to improve this method when the missingness mechanism is nonignorable *(NMAR)*, we create an additional dummy Boolean variable R_j (missingness indicator), $j = 1, 2, ..., k$, to represent missingness for each existing variable X_j that was found to be incompletely observed. Actually, for each case in the incomplete variable X_j, the variable R_j takes on one of the two values 1 and 0, denoting respectively, that the entry X_j is observed or not. Therefore we have a dataset consisting of the original data set D = (O,M) and the additional dummy variables R, D_R = (O,M,R). In this improved method, learning the Bayesian network from data has some structural constraints. It means that for all j in $\{1, 2, \ldots, k\}$, there must be an edge from the R_j node to the related X_j node. The graphical structure of the Bayesian network representing the joint probability distribution of the variables in the D_R = (O,M,R) can be used to help improve the performance of BN method when the missingness mechanism is IM. Our approach is based on the imputation method described in section 3.2, but the visual inspection of the graph induced from the augmented data set, D_R, will account for the missing pattern, and hence will improve the performance of the Bayesian network in imputation (BN_R method).

Performance evaluation

The main purpose of using Bayesian networks for imputation is to preserve the variables' relationship as much as possible. To see whether the mentioned goal is achieved, we check the joint distribution. This joint distribution checking is entitled *statistical consistency* by Di Zio *et. al.* (2004).

To evaluate the preservation of distributions, two indexes are applied. A delta index which is described in the next subsection and a modified version of the index *Kullback-Leibler information* (Kullback & Leibler 1951; Kullback 1959; Kullback 1987) called *Minimum Kullback-Leibler information* which is outlined in subsection entitled: '*Minimum Kullback-Leibler index'*.

Delta index

Beginning with a complete data set (i.e. a single variable X with total number of n records), missing values of size n^* were artificially produced. The relative frequency of category x of X in the original data set, is denoted by f_x, and the frequency of the same category x of X after imputation is denoted by $\tilde{f}_x$. A distance function that can be defined for these two frequencies is:

$$\Delta = \frac{1}{2}\sum_{X} |f_X - \tilde{f}_X|, \tag{2}$$

Where the sum is over all categories of X. The above index takes values between 0 and 1. This index can be extended easily to a multivariate context in order to check the preservation of joint distributions. For example, consider two variables X_i and X_j. Then the delta-indicator assumes the form:

$$\Delta = \frac{1}{2}\sum_{x_i}\sum_{x_j} |f_{x_i x_j} - \tilde{f}_{x_i x_j}|. \tag{3}$$

Minimum Kullback-Leibler index

One way of studying the effect of imputed missing items on the distribution of data (and hence on the joint relationship of variables) is to measure the distance between two distributions, which are the above explained f_x and $\tilde{f}_x$.

Suppose that T is a random variable and f and g are two densities, then the Kullback-Leibler distance of g to f is defined as:

$$K[f, g] = \int \log[\frac{f(t)}{g(t)}]f(t)dt. \tag{4}$$

It is crucial to know how much information is lost. The wider the distance between two distributions, the more information we have lost by imputation. The logarithm is a transformation which maps small distances to wider ones. So taking the logarithms of f and g, and subtracting them helps to better observe the small differences between these two probability distributions. This feature of the logarithm function leads us to detect better the subtle changes in joint relationships of variables that may arise by imputation. Again suppose that n^* is the number of artificially produced missing values. f_x and $\tilde{f}_x$ are also defined as above. The adjusted formula for the two discrete distributions is:

$$K[\tilde{f}_x, f_x] = \sum \log[\frac{\tilde{f}_x}{f_x}]\tilde{f}_x, \tag{5}$$

With the assumption that $0 \times \log(0) = 0$. The Kullback-Leibler information is not symmetric in f and g, in other words, $K[f, g] \neq K[g, f]$.

A modification, remedies this situation. The minimum of $K[f,g]$ and $K[g,f]$, denoted by MKL[f, g] and is formulated as follows:

$$MKL[f,g] = \min(K[f,g], K[g,f]). \tag{6}$$

This helps us to have a symmetric index $MKL[f,g]$ to measure the distance between two distributions f and g.

INFORMATION EVALUATION

In this paper, Bayesian networks are used for missing data imputation in the presence of three different missingness mechanisms (Missing At Random, Missing Completely At Random, Missing Not At Random). But the important question is 'Will the performance of Bayesian networks be affected by different missingness mechanisms?' The question is answered using the evaluation methods mentioned in subsection 3.3. Use of Entropy and Mutual Information, also helps to assess better the effect of missingness mechanisms on the performance of Bayesian networks in the imputation of missing data.

Entropy is a measure of how much the probability mass is scattered around on the states. In fact, the more random a variable is, the higher its entropy will be. Let X be a discrete random variable with n states $x_1, x_2, ..., x_n$ and probability distribution $P(X)$, then the entropy of X is defined as:

$$H(X) = -\sum_{i=1}^{n} P(X = x_i) \times \log P(X = x_i), \tag{7}$$

which is greater than zero. If Y is another random variable, then the mutual information of variables X and Y is computed as:

$$I(X, Y) = \sum_{Y} P(Y) \sum_{X} P(X|Y) \log \frac{P(X, Y)}{P(X)P(Y)} \tag{8}$$

The mutual information, I(X,Y), is a measure of the information shared by X and Y. If X is the variable of primary interest, then I(X,Y) is a measure of the value of observing Y.

Value of information analysis

Value of information analysis is the task of identifying the values of pieces of information. Considering a Bayesian network, entropy and mutual information can be used in defining a value function.

Suppose X is the variable to be analyzed, in order to keep the idea to let high values be preferred, we let an entropy-based value function be (Jensen & Jianming 1995):

$$V(X) = -H(X) \tag{9}$$

Also the value of the information after observing a variable Y is:

$$V(X|Y) = -(H(X) - I(X, Y)) \tag{10}$$

The reason for computing these values is to identify the variable, which increases the value of information the most. In order to examine the effect of the missing data mechanism, we have defined a Mean Value of Information (MVI) index. Consider a Bayesian network with nodes $\{X, Y_1, ..., Y_n\}$, the MVI index for variable X is defined as:

$$MVI(X) = \frac{\sum_{i=1}^{n} V(X|Y_i)}{n}. \tag{11}$$

This means value of information changes as the pattern of missingness varies between ignorable ones to non-ignorable ones, and will be most affected by latter mechanism.

EXPERIMENTAL RESULTS

This section is devoted to apply the Bayesian network based imputation methods described in Sections 3.2 and 3.3, on two data sets. The first data set is an experimental study on a data set of individual records obtained from the Iran Statistical Research Centre, which is about the urban families Household Income and Expenditure (HIE). The second data set is extracted from the study that examines a sample of 405 children who are within the first two years of entry to elementary school. We use HUGIN software (www.hugin.com) and R (http://cran.r-project.org/) for the analysis of these data.

Description of the urban families HIE data set

Data set of urban families HIE was obtained from the 2005 Iran Statistical Center census. According to this data set, the seven variables are:

- Sex: man (1) and woman (0);
- Poverty: absolute poverty (1), quasi poor (2), non-poor (3) and quasi rich or rich (4)

- Educational level: literate (1) and illiterate (0);
- Occupation status: employed (1), non-employed (seeking a job) (2), students (3) and others (4);
- Marital status: married (1), widow(er) (2), divorced (3) and single (4);
- Residential status: proprietor (1), tenant (2) and others(3);
- Family size: one member (1), two members (2),. . ., six members (6).

Description of the antisocial behavior in children data

This data set was collected using face-to-face interviews of both the child and the mother taken at two-year intervals between 1986 and 1992. The study examines the measures of emotional support and cognitive stimulation provided to the child by the mother near the time of school entry and also examines four repeated measures of both the child's antisocial behavior and the child's reading recognition skills (http://www.duke.edu/~curran/).

Here, applying some preliminary analysis on the original data set, the following three variables collected at year 1986 were chosen to be worked on:

- Anti: child's antisocial behavior, not true (0), sometimes true (1) or often true (2)
- Homecog: child's cognitive stimulation at home, if the family gets a daily newspaper (1), if the family encourages their child to start and keep doing hobbies (2), if the mother often reads stories to her child (3)
- Homemo: child's emotional support at home, the mother encouraged the child to contribute to the Conversation (1), the mother answered the child's questions or requests verbally (2), the mother's voice conveyed positive feelings about the child (3).

Application of methods and results

Let us use the following abbreviations for the variables of the first data set: 1. Sex: sex, 2. Marital status: mar, 3. Occupation status: occ, 4. Residential status: acc, 5. Educational level: edu, 6. Family size: fsize and 7. Poverty level: pov. These abbreviations are used in the estimated Bayesian network of the families HIE data, to indicate the nodes' names (see Fig. 1.(a)). Also in Fig. 1(b) you can find the Bayesian network constructed from the antisocial behavior data set.

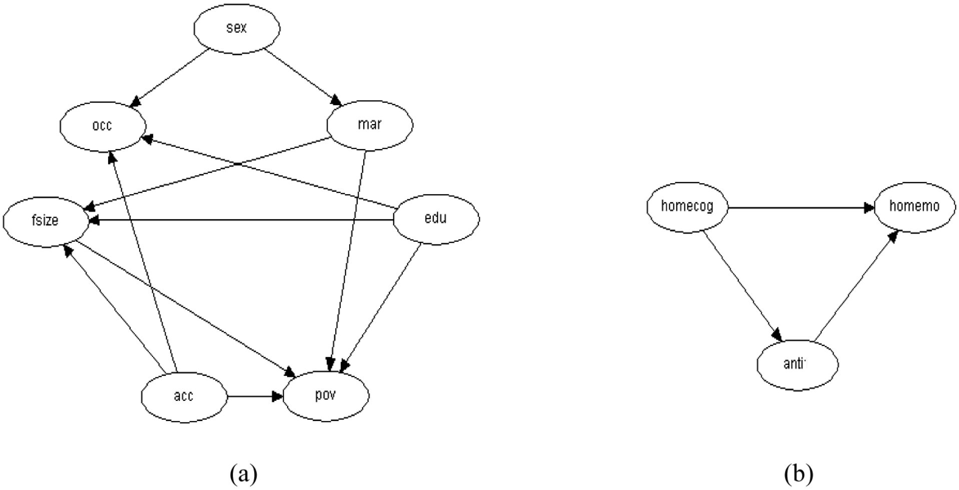

Fig.1. (a) The estimated Bayesian network from the complete data set of 'urban families HIE' data and (b) The estimated Bayesian network from the 'antisocial behavior in children' data.

To implement the imputation method based on Bayesian networks, following two steps should be done. At first we randomly generate artificial missing items in the set of complete sample data, then the Bayesian network is learnt from the perturbed data set, D = (O,M), and finally it is used for imputing missing items (BN method).

To apply the imputation method based on augmented Bayesian networks, after artificial perturbation of the data, the Indicator nodes (R) should be added to the perturbed data set to obtain the D_R = (O,M,R). Then the Bayesian network is learnt from D_R and used for the imputation process (BN_R method).

For the urban families HIE data set, just the poverty node is contaminated. But in the antisocial behavior data, missing items are generated in all nodes. So we have the augmented data sets D_R = (O = (sex,occ,mar,edu,fsize,acc), M = (pov), R = (Rpov) and D_R = (M = (anti, homecog, homemo), R = (Ranti, Rhomecog, Rhomemo)) for urban families HIE and antisocial behavior data sets, respectively.

Applying the above steps, three experiments were carried out. In the first experiment, only missing completely at random (MCAR) items are generated. In the second, missing at random (MAR) items are generated and in the third experiment, not missing at random (NMAR) items are generated.

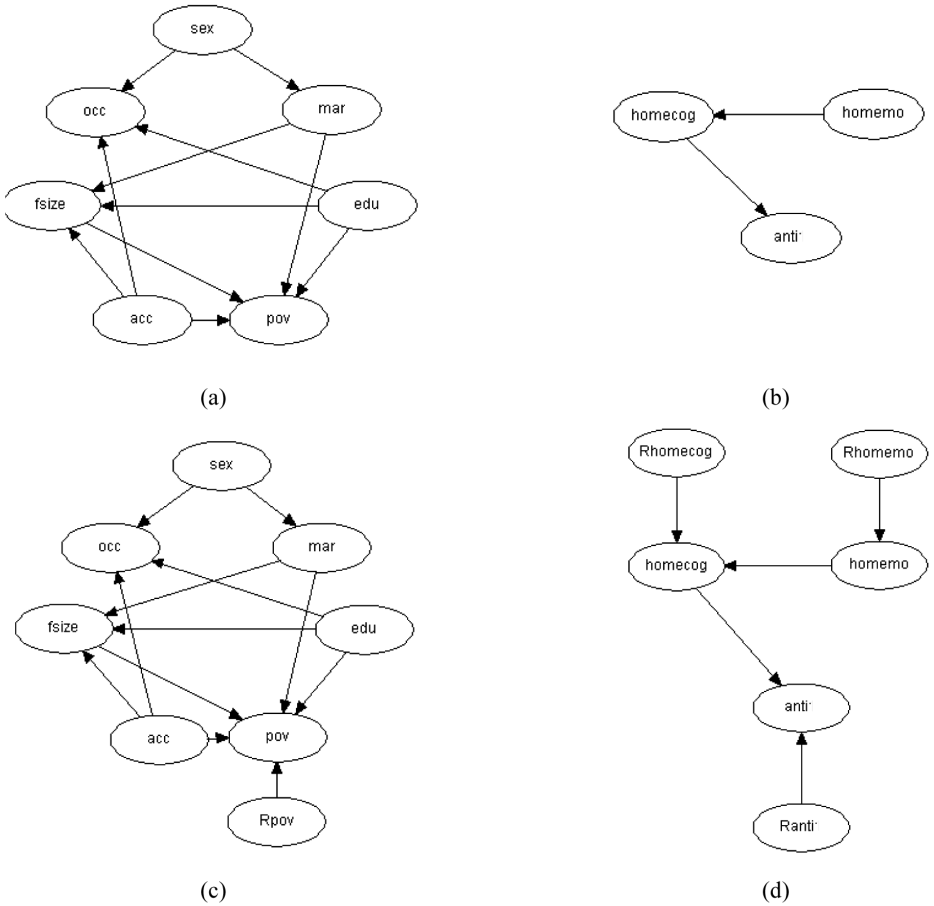

Fig.2. Bayesian networks in parts (a) and (b) are based on the perturbed data sets D = (O,M), according to the MCAR mechanism, Bayesian networks in parts (c) and (d) are based on the augmented data sets D_R = (O,M,R) and according to the MCAR mechanism (R_* is the observed indicator variable for variable $_*$, e.g. Rpov is the missing indicator variable for the pov node)

Figure 2 is devoted to the Bayesian networks constructed from data sets D and D_R in the presence of MCAR mechanism, D as in (a) for urban families HIE and (b) for antisocial behavior data, and D_R as in (c) for the urban families HIE and (d) for antisocial behavior data.

The aim of this part is to evaluate the effects of different missingness mechanisms on the performance of the BN and BN_R method and to compare it with that of hot-deck methods using two indices, delta and minimum Kullback-Leibler.

According to the definition of ‘reliability order’, and having contaminated just the poverty node of the urban families HIE data, it is not allowed to have any edges from poverty to the other variables. For the antisocial behavior data the whole data set is contaminated and it also obeyed the ‘reliability order’ rule.

Benchmarking the BN and BN_R imputation method, two kinds of hot-deck imputation were carried out: random overall hot-deck imputation and within-class random hot-deck imputation. Random overall hot-deck imputation is a technique where a respondent is chosen at random from the total respondent sample, and the selected respondent’s value is assigned to the non-respondent. The second method differs in that a donor respondent is chosen at random within the same class as the recipient non-respondent.

Once the qualitative and the quantitative parts of the Bayesian network have been estimated, each missing item can be imputed according to the BN based methodology, BN_R based method and the above hot-deck imputation methods. The Mean Square Error (MSE) of delta (Δ) and Minimum Kullback-Leibler (MKL) indices, have been approximated through a Monte Carlo experiment consisting of $t = 1, ..., 1000$ replications of the Bayesian network and hot-deck imputation algorithms on the perturbed data set. The approximations are:

$$MSE(\Delta) = E(\Delta - \tilde{\Delta})^2 = E(\Delta - \bar{\Delta})^2 + (\bar{\Delta} - \tilde{\Delta})^2 = Var(\Delta) + \bar{\Delta}^2, \tag{13}$$

$$\bar{\Delta} = \frac{1}{1000}\sum_{t=1}^{1000} \Delta_t, \qquad Var(\Delta) = \frac{1}{1000-1}\sum_{t=1}^{1000}(\Delta_t - \bar{\Delta})^2$$

where $\bar{\Delta}$ is the real value of the Δ and since,

$$\forall t \in \{1, ..., 1000\}, \qquad \tilde{\Delta}_t = \frac{1}{2}\sum_x |f_x - f_x| = 0. \tag{14}$$

And:

$$\begin{aligned} MSE(MKL) \quad &= E(MKL[f_x, \tilde{f}_x] - \tilde{M}KL[f_x, \tilde{f}_x])^2 \\ &= E(MKL[f_x, \tilde{f}_x] - \bar{M}KL[f_x, \tilde{f}_x])^2 \\ &+(\bar{M}KL[f_x, \tilde{f}_x] - \tilde{M}KL[f_x, \tilde{f}_x])^2 \\ &= Var(MKL[f_x, \tilde{f}_x]) + \bar{M}KL^2[f_x, \tilde{f}_x], \end{aligned} \tag{15}$$

$$\bar{M}KL[f_x, \tilde{f}_x] = \frac{1}{1000}\sum_{t=1}^{1000} J_t[f_x, \tilde{f}_x],$$

$$Var(MKL[f_x, \tilde{f}_x]) = \frac{1}{1000-1}\sum_{t=1}^{1000}(MKL_t[f_x, \tilde{f}_x] - \bar{M}KL[f_x, \tilde{f}_x])^2$$

where again $\tilde{M}KL[f_x,\tilde{f}_x] = MKL[f_x,f_x] = \frac{1}{1000}\sum_{t=1}^{1000} MKL_t[f_x,f_x] = 0$ is the real value of the $MKL[f_x,\tilde{f}_x]$, since for all $t \in \{1, ..., 1000\}$:

$$MKL_t[f_x,f_x] = \min(K_t[f_x,f_x], K_t[f_x,f_x]) = K_t[f_x,f_x] = \sum \log[\frac{f_x}{f_x}]f_x = 0. \quad (16)$$

The results of the urban families HIE data set and the antisocial behavior in children is reported, respectively, in Tables 2 and 3. These results indicate the high efficiency of Bayesian network-based method compared to that of the augmented Bayesian network method and those of other hot-deck methods, in the field of imputation.

Table 2. Mean Square Error (MSE) of Minimum Kullback-Leibler (MKL) index and Delta (Δ) index for the Bayesian Network (BN), augmented Bayesian Network (BN$_R$), stratified hot-deck (HD str) and random overall hot-deck (HD) procedures considering different missing mechanisms for the urban families HIE

Method		Mechanism		
		MCAR	MAR	NMAR
BN	MSE(Δ)	0.027443	0.023547	0.172683
	MSE(MKL)	0.005112	0.003399	0.248500
BN$_R$	MSE(Δ)	0.045512	0.030052	0.057911
	MSE(MKL)	0.006759	0.002639	0.007838
HD (str sex)	MSE(Δ)	0.214174	0.244689	0.283522
	MSE(MKL)	0.199882	0.258929	0.586604
HD (str mar)	MSE(Δ)	0.212764	0.236859	0.281741
	MSE(MKL)	0.206698	0.240846	0.585196
HD (stracc)	MSE(Δ)	0.195537	0.211987	0.246394
	MSE(MKL)	0.181578	0.207850	0.466240
HD (str sex-mar)	MSE(Δ)	0.211724	0.227452	0.273990
	MSE(MKL)	0.199471	0.225150	0.497904
HD	MSE(Δ)	0.234104	0.160303	0.320000
	MSE(MKL)	0.235156	0.270339	0.640070

Table 3. Mean Square Error (MSE) of Minimum Kullback-Leibler (MKL) index and Delta (Δ) index for the Bayesian Network (BN), augmented Bayesian Network (BN_R), stratified hot-deck (HD str) and random overall hot-deck (HD) procedures considering different missing mechanisms for the antisocial behavior data set

Method		Mechanism		
		MCAR	MAR	NMAR
BN	MSE(Δ)	0.243170	0.194110	0.619878
	MSE(MKL)	0.504991	0.497553	3.486397
BN_R	MSE(Δ)	0.357831	0.282882	0.467079
	MSE(MKL)	0.546900	0.542372	1.447663
HD str	MSE(Δ)	0.377957	0.342103	0.682530
	MSE(MKL)	1.085293	1.288636	5.303826
HD	MSE(Δ)	0.392962	0.348107	0.685602
	MSE(MKL)	1.403290	1.963572	5.319799

As far as the MAR and MCAR mechanisms are concerned, both indices indicate preferences for Bayesian network instead of augmented Bayesian network and hot-deck methods, although the BN and BN_Rmethods are comparable. It can also be concluded that the small (large) distances between distributions of the original (f_x) and the imputed ($\tilde{f}_x$) data sets, are mapped in the smaller (larger) distances by the Minimum Kullback-Leibler index. This feature of the Minimum Kullback-Leibler index, makes the differences more prominent.

Generally Bayesian network method gives more acceptable results when the missingness mechanism is MAR or MCAR showing the sensitivity of Bayesian networks to the missingness mechanism. In NMAR, in contrast to the MAR and MCAR mechanisms, there is no such improvement in indices for Bayesian network based imputation. But we observe a significant increase in the efficiency of the augmented Bayesian network method relative to that of Bayesian network method and those of hot-deck imputation methods, when the mechanism is IM.

Tables 4 and 5 are devoted, respectively, to the results of the value of information analysis of the urban families HIE data set and that of the antisocial behavior of the children data set. From Table 4, we can conclude that in all missing patterns, the most informative node for poverty level is the family size. But in the Bayesian network constructed from the contaminated data by the NMAR mechanism in the poverty node, a decrease of mutual information of ‘fsize’, ‘acc’ and ‘edu’ with ‘pov’ (which is not desired) has been compensated by a decrease in the entropy of ‘pov’ (which is seemingly desired) and results in a significant increase in MVI index for ‘pov’ (from -1.3432 for the complete data set to -1.2283). It is because of the nonrandom contamination in categories of the poverty variable, which leads to a decrease in uncertainty (entropy) of the variable.

Table 4. Value of information analysis for poverty node of the urban families HIE data set; H(.): entropy, I(.,.): mutual information and MVI(.): Mean Value of Information

	Mechanisms			
	Complete data	MCAR	MAR	NMAR
H(pov)	1.3600	1.3600	1.3500	1.2400
I(fsize, pov)	0.0800	0.0900	0.1000	0.0700
I(acc, pov)	0.0200	0.0200	0.0000	0.0000
I(edu, pov)	0.0004	0.0006	0.0000	0.0000
I(mar, pov)	0.0000	0.0000	0.0000	0.0000
I(occ, pov)	0.0000	0.0000	0.0000	0.0000
I(sex, pov)	0.0000	0.0000	0.0000	0.0000
MVI(pov)	-1.3432	-1.3415	-1.3333	-1.2283

In Table 5, when the missing pattern is NMAR, again we can see an obvious decrease in the entropy and the mutual information for all nodes, which leads to a significant increase in the MVI indices of the nodes. One of the consequences of this seemingly good event (the increase in MVI) is the wrong decisions which might be taken on the basis of this increased MVI. For the other two mechanisms, the MVI index is approximately the same as the one for the complete data.

Table 5. Value of information analysis for all nodes of antisocial behavior; H(.): entropy, I(.,.): mutual information and MVI(.): Mean Value of Information

	Mechanisms			
	Complete data	MCAR	MAR	NMAR
H(anti)	0.7300	0.7400	0.7300	0.6700
H(homecog)	0.9000	0.9000	0.9000	0.6900
H(homemo)	0.8700	0.8500	0.8700	0.7400
I(anti, homecog)	0.0300	0.0400	0.0500	0.0000
I(anti, homemo)	0.0200	0.0009	0.0200	0.0300
I(homemo, homecog)	0.0500	0.0500	0.0500	0.0000
MVI(anti)	-0.7050	-0.7190	-0.6950	-0.6550
MVI(homecog)	-0.8600	-0.8550	-0.8500	-0.6900
MVI(homemo)	-0.8350	-0.8240	-0.8350	-0.7250

CONCLUSION

To assess the benefits and drawbacks of the BN_R approach, an experimental study was conducted on two data sets in order to compare the introduced method, BN_R, with other three methods of missing data imputation methods: random overall hot-deck imputation, within-class random hot-deck imputation and Bayesian network based. Compared to the Bayesian network based imputation method (BN method) of Di Zio *et. al.* (2004) and other traditional imputation methods, BN_Rmethod performed much better in the presence of NMAR mechanism. But in ignorable mechanisms although BN_R method performs better than hot-deck methods, BN method is still preferable.

Imputation by means of Bayesian networks seems to be a reasonable approach to improve the consistency of imputed data sets (BN for ignorable and BN_R for nonignorable mechanisms). This conclusion was reached because of the good results obtained in terms of the preservation of joint distributions. The preservation of joint distributions was evaluated by means of the delta index defined in the literature and the Minimum Kullback-Leibler index introduced here for the probability distribution distance. As can be seen from Tables 2 and 3, the latter index is more convenient for comparing small differences. Besides, the use of Bayesian networks when the contamination mechanism is nonignorable causes the value of information, which in this study is measured by the MVI index, to wrongly increase the entropy.

REFERENCES

Aussem A., & Rodrigues de Morais S. 2010. A Conservative feature subset selection algorithm with missing data. Neurocomputing, **73**: 585-590.

Buntine, W. L. 1994. Operations for learning with graphical models. Journal of Artificial Intelligence Research **2**: 159-225.

Cowell, R. G., Dawid, A. P., Lauritzen, S. L. & Spiegelhalter, D. J. 1999. Probabilistic networks and expert systems. Springer, New York.

Di Zio, M., Scanu, M., Coppola, L., Luzi, O. & Ponti, A. 2004. Bayesian networks for imputation. Journal of Royal Statistical Society,A,Vol. **167**(2): 309-322.

Di Zio, M., Sacco G., Scanu M. & Vicard P. 2005. Multivariate techniques for imputation based on Bayesian networks. Neural Networkworld .**4**:.303-309.

Heckerman, D. 1996. A tutorial on learning with Bayesian networks. Microsoft Research Technical report, MSR-TR-95-06.

Jensen, F. V. 1996. An introduction to Bayesian networks, Springer-Verlag, New York.

Jensen, F. V. & Jianming, L. 1995. dr-Hugin: A System for hypothesis driven data request. I **Gammerman, A.** (ed.) (red.), Probabilistic Reasoning and Bayesian Belief Networks.

Jensen, Finn V. & Nielsen Thomas D. 2007. Bayesian networks and decision graphs, Springer, New York.

Kullback, S. & Leibler, R.A. 1951. On information and sufficiency. Annals of Mathematical Statistics **22**(1): 79--86.

Kullback, S. 1959. Information theory and statistics, JohnWiley and Sons, New York.

Kullback, S. 1987. Letter to the Editor: The Kullback-Leibler distance. The American Statistician **41**(4):.340--341.

Lauritzen, S. 1995. The EM algorithm for graphical association models with missing data. Computational Statistics and Data Analysis **19** 191-201.

Lauritzen, S. L. & Spiegelhalter, D. J. 1988. Local computations with probabilities on graphical structures and their application to expert systems (with discussion). Journal of the Royal Statistical Society, Series B **50**:157-224.

Lin J. & Haug P. 2008. Exploiting missing clinical data in Bayesian network modeling for predicting medical problems. Journal of Biomedical Informatics **41**: 1-14.

Pearl, J. 1988. Probabilistic reasoning in intelligence systems, Morgan Kaufmann, San Mateo, California.

Ramoni, M. & Sebastiani P. 2001. Robust learning with missing data. Machine Learning **45**:147-170.

Rodrigues de Morais, S. & Aussem, A. 2009. Exploiting data missingness in Bayesian network modeling. In the proceedings of 8th International Symposium on Intelligent Data Analysis (IDA 2009), LNCS N°5772, Springer-Verlag, Lyon, France 35-46.

Rubin, D. B. 1976. Inference and missing data. Biometrika, **63**: 581-592.

Spirtes, P., Glymour, C. & Scheines, R. 1993. Causation prediction and search, Springer-verlag, New York.

Thibadeau, Y. & Winkler, W. E. 2002. Bayesian networks representations, generalizied imputation and synthetic micro-data satisfying analytic constraints", Technical report RRS2002/9, U.S. Bureau of the Census.

Submitted : *1/12/2011*
Revised : *12/6/2012*
Accepted : *18/6/2012*

تحسين أداء شبكات بايزي في عدم تجاهله احتساب البيانات المفقودة

[*]ب. نيلوفار، [**]م. جانجالي، [***]م. ر.فاريد روحان

قسم الاحصاء - كلية علوم الرياضيات - جامعة شهيد بهشتي - إيران

خلاصة

فقدان البيانات من الامور التي قد تواجه الباحثين الذين يتعاملون مع مسائل جمع البيانات. وتعتبر شبكات بايزي هي أحدى الطرق المقترحة التي يتم استخدامها مؤخراً في احتساب البيانات المفقودة. الهدف الرئيسي من هذا البحث هو تحسين كفاءة شبكات بايزي في عدم تجاهله إحتساب البيانات المفقودة من خلال إضافة روابط المؤشر المفقودة للمتغيرات غير المكتملة وبناء شبكة بايزي مطورة. وكذلك النظر في تأثير أنواع مختلفة من آلية الفقدان (التجاهليه وعدم التجاهليه) على أداء طرق الإسناد. وقد تم مقارنة أربع طرق إسناد مختلفة وهي: عموما عشوائية ساخنة السطح، ضمن فئه عشوائية ساخنة السطح، باستخدام شبكات بايزي وباستخدام شبكات بايزي مطورة، وذلك باستخدام رقمين قياسيين: (1) دالة مسافة و(2) مؤشر" كولباك ليبلير" للحد الادنى. وتشير النتائج إلى الجودة العالية في هذه الطرق استناداً إلى شبكات بايزي مقارنة بغيرها من طرق الاسناد.

المجلــة التربــوية

مجلة فصلية، تخصصية، محكّمة

تصدر عن مجلس النشر العلمي - جامعة الكويت

رئيس التحرير: أ. د. عبدالله محمد الشيخ

المجلة التربوية
مجلة فصلية، تخصصية، محكمة
تصدر عن مجلس النشر العلمي - جامعة الكويت
بحوث
مراجعات
محاضرات
تقارير
ملخصات

تنشر:

- البحــوث التربــوية المحكمة
- مراجـعات الكتب التربوية الحديثة
- محاضــر الحوار التربوي
- التقارير عــن المؤتمـرات التربـوية وملخصـــات الرســـائل الجـــامعية

- تقبل البحوث باللغتين العربية والإنجليزية.
- تنشر لأساتذة التربية والمختصين بها من مختلف الأقطار العربية والدول الأجنبية.

الاشتراكات:

في الكويـــــــت: ثلاثة دنانير للأفـراد، وخمسة عشر ديناراً للمؤسسات.

في الدول العربيـــة: أربعـة دنـانير للأفراد، وخمسة عشر ديناراً للمؤسسات.

في الدول الأجنبية: خمسة عشـر دولاراً للأفراد، وستون دولاراً للمؤسسات.

توجه جميع المراسلات إلى:

رئيس تحرير المجلة التربوية - مجلس النشر العلمي ص.ب. ١٣٤١١ كيفان - الرمز البريدي 71955
الكويت هاتف: ٢٤٨٤٦٨٤٣ (داخلي ٤٤٠٣ - ٤٤٠٩) - مباشر: ٢٤٨٤٧٩٦١ - فاكس: ٢٤٨٣٧٧٩٤
E-mail: joe@ku.edu.kw