Expert Systems with Applications 38 (2011) 442–449

Contents lists available at ScienceDirect

Expert Systems with Applications

journal homepage: www.elsevier.com/locate/eswa

Diagnose the mild cognitive impairment by constructing Bayesian network with missing data

Yan Sun [a,b,c], Yiyuan Tang [a,*], Shuxue Ding [d], Shipin Lv [c], Yifen Cui [a]

[a] Neuroinformatics Institute, Dalian University of Technology, Dalian 116024, China
[b] Department of Computer Science, Liaoning Normal University, Dalian 116029, China
[c] Department of Computer Science and Engineering, Dalian University of Technology, Dalian 116024, China
[d] School of Computer Science and Engineering, Aizu University, Aizu-Wakamatsu City, Fukushima 965-8580, Japan

ARTICLE INFO

Keywords:
Mild cognitive impairment (MCI)
Missing data
Bayesian network
Mutual information
Newton interpolation

ABSTRACT

Mild Cognitive Impairment (MCI) is thought to be the prodromal phase to Alzheimer's disease (AD), which is the most common form of dementia and leads to irreversible neurogenerative damage of the brain. In order to further improve the diagnostic quality of the MCI, we developed a MCI expert system to address MCI's prediction and inference question, consequently, assist the diagnosis of doctor. In this system, we mainly deal with following problems: (1) Estimate missing data in the experiment by utilizing mutual information and Newton interpolation. (2) Make certain the prior feature ordering in constructing Bayesian network. (3) Construct the Bayesian network (We term the algorithm as MNBN). The experimental results indicate that MNBN algorithm achieved better results than some existing methods in most instances. The mean square error comes to 0.0173 in the MCI experiment. Our results shed light on the potential application in MCI diagnosis.

1. Introduction

Alzheimer's disease (AD) is the most common form of dementia and that may lead to irreversible neurogenerative damage of the brain. But the current diagnostic tools have poor sensitivity, especially for the early stages of AD and are not easy to be diagnosed until AD has led to irreversible brain damage (Morris et al., 2001). Therefore, it is very important research topic for how to diagnose AD as early as possible. Through research effort of recent 10 years, it is concluded that MCI (Mild cognitive impairment) is the early stage of the Alzheimer Diseases (Celsis, 2000; Morris et al., 2001; Petersen et al., 2001). 10–30% of MCI patients convert to AD annually, whereas the rate of conversion of cognitively normal elderly people is 1–2% (Celsis, 2000). Furthermore, there is evidence that 100% of patients with MCI progress to greater dementia severity (Petersen et al., 2001). So the problem of diagnosing AD can be converted into the diagnosis of the MCI. Up to now, however, there is still not a strict and unified standard.

In this study, we develop a specific diagnostic system on the MCI, which predicts and diagnoses the MCI by using some artificial intelligent methods. Since a practical database usually might be not complete, at first, we utilize the mutual information and Newton interpolation to estimate the values of missing data. Then, we propose to determine the feature ordering by using the mutual information and defining a "higher filter". Finally, we construct the Bayesian network for assisting the prediction and diagnosis of the MCI.

The remainder of this paper is organized as follows: Section 2 briefly reviews some related works. In Section 3, we present the MNBN algorithm. In Section 4, we further describe how to implement the MNBN algorithm. In Section 5, we report and analyze experimental results. Finally, in Section 6, we draw the main conclusions and give some discussions.

2. Related works

In recent years, one new idea is to assistant diagnose the MCI by using some method of artificial intelligence. Among them, Bayesian Network is popular within the community of artificial intelligence due to their ability to support probabilistic reasoning from data with uncertainty. According to the network, probabilistic inference can be conducted to predict the values of some variables based on the observed values of other variables. Hence, Bayesian networks are widely used in many areas. Reference (Chen & Herskovits, 2006) applied Bayesian Network to model the interactions among morphological changes and clinical variables of the MCI. We can conclude four principal advantages of using a discrete variable Bayesian network for network analysis:

* Corresponding author. Address: Neuroinformatics Institute, Dalian University of Technology, Dalian, No. 2 Linggong Road, Ganjingzi District, Dalian City, Liaoning Province 116024, China. Tel.: +86 411 84706039; fax: +86 411 84706046.
E-mail address: yiyuan@uoregon.edu (Y. Tang).

0957-4174/$ - see front matter © 2010 Elsevier Ltd. All rights reserved.
doi:10.1016/j.eswa.2010.06.084

(1) the Bayesian network framework does not require that the joint distribution follows a specific parametric distribution; (2) a Bayesian network supports probabilistic reasoning as it consists of probabilistic associations among variables; (3) because the Bayesian network representation is based on the concept of conditional independence, it supports Bayesian inference without having to maintain the full joint distribution in memory; and (4) since each Bayesian network is a multivariate model that we can evaluate using a single probability score, we can evaluate many structure–function interactions without the multiple comparison problem.

Although the method of network analysis is very effective, such as the method of reference (Wong & Leung, 2004; Chen & Herskovits, 2006; Liang & Zhang, 2009), these algorithms can not deal with the missing data. Whereas, in real-world applications, missing data is a great quantity, especially in the medical problem. In the MCI experiment, because the eyesight of some subjects is not better, they can not see clearly the stimulation, which leads to appear some missing values in the data set. One example of the importance of handling missing data is that more than 40% of data sets in the UCI repository have missing values (Newman, Hettich, & Blake, 1998; Garcia-Laencina, Sancho-Gomez, Figueiras-Vidal, & Verleysen, 2009), which is one of most commonly used data sets for benchmarking machine learning procedures. Many high efficient and effective learning algorithms require complete data sets to execute. The conversion from an incomplete data set to a complete one then becomes an issue.

Many researchers have been working on constructing Bayesian network from incomplete data sets. However, there are a few algorithms available for learning the Bayesian network structure with missing data, because most algorithms require a complete data set (Lin & Haug, 2008). For learning Bayesian network from incomplete data set, the most important challenge is that the parameter values and the scores of networks can not be computed directly on the cases with missing values. Moreover, the scoring metric can not be decomposed directly. Thus, a local change in the network structure will lead to reevaluate the score of the whole network (Wong & Guo, 2008).

Friedman proposes a Bayesian Structural Expectation Maximization (SEM) algorithm which alternates between the parameter optimization process and the model search process (Friedman, 1997, 1998). However the algorithm takes much longer time to run and is lack of stability (Austin & Escobar, 2005; Lin and Haug, 2008). The score of a Bayesian network is maximized by means of the maximization of the expected score. Pena et al. uses the BC + EM method instead of the EM method (Dempster, Laird, & Rubin, 1977) in their BS-BC + EM algorithm for clustering (Pena, Lozano, & Larranaga, 2000; Pena, Lozano, & Larranaga, 2002). However, the search strategies adopted in most existing SEM algorithms may not be effective and may make the algorithms find suboptimal solutions. Myers et al. employ a genetic algorithm to learn Bayesian networks from incomplete data sets (Myers, Laskey, & DeJong, 1999). Both network structures and the missing values are encoded and evolved. The incomplete data set is completed by specific genetic operators during evolution. Nevertheless, it has the efficiency and convergence problems because of the enlarged search space and the strong randomness of the genetic operators for completing the missing values.

It is worth mentioning that Wong uses evolutionary algorithm to learn Bayesian network from incomplete data sets, called EBN (Evolutionary Bayesian Network learning method) (Wong & Guo, 2008), which utilizes the efficient and effective global search ability of HEA (Wong & Leung, 2004) and applies EM (Dempster et al., 1977) to handle missing values. However, EBN is a stochastic algorithm and results are strongly dependent on the initial network structure, so the results are not stationary.

In this paper, we propose a novel method that firstly uses mutual information to get the important extent of the feature. According to the importance of feature we find the most similar cases with the missing value case. Then, we adopt the Newton interpolation to estimate the value of the missing data. Finally, we construct the Bayesian network by using K2 algorithm (Cooper & Herskovits, 1992), it is the most effective, efficient and most popular. However, this algorithm has one disadvantage that it must specify a prior feature sequence. The feature ordering consists of domain knowledge or constraints that specify a partial order, such that a parent feature must appear earlier in the order than any of its descendants (Chen & Herskovits, 2006). The prior feature ordering most depends on the subjective experience of researchers, which serious effect on the results of the Bayesian network model. Some researcher used the oriented tree obtained from the maximum-weight spanning-tree algorithm to generate this ordering (Heckerman, Geiger, & Chickering, 1994; Chen & Herskovits, 2006). In this study, we utilize the mutual information and define a "higher filter" to learn the prior feature ordering.

3. The algorithm

We suppose that there is a data set D(or sampling space) with $X = \{x_1, \ldots, x_n\} \subset R^d$, for each case $x_a \in X(a = 1, \ldots, n)$ has m features $F = \{f_1, \ldots f_m\}$, it can be represented as a value vector of features, i.e., $x_a = (v_{a1}, \ldots, v_{am})$, where v_{ai} is the value of x_a corresponding to the feature f_i. Among them some v_{ai} are missing and the number of missing data is k.

Given a data set D, the first objective of learning algorithm is to get the estimation of the missing data v_{ai} by computing the mutual information and Newton interpolation. The second objective is to get the prior feature ordering by defining a "higher filter". The third goal is to construct the Bayesian network B_s using K2 algorithm (Cooper & Herskovits, 1992), it will find the nonlinear relationships among all the features and get the posterior distribution for the functional feature MCI from the Bayesian network. That is, we can predict the state of feature MCI based on the Bayesian network.

3.1. Getting the relationships between features

Since mutual information is good at quantifying how much information is shared by two random variables, it is often taken as evaluation criterion to measure the relevance between features and the class labels (Marcus, Hutter, & Zaffalon, 2005; Liu, Sun, Liu, & Zhang, 2009). In this study, we utilize mutual information to measure the relationship between features, which aims to estimate the missing feature data.

We assume that $f_i \subseteq F$ and $f_j \subseteq F(1 \leqslant i,j \leqslant m)$ represent the selected missing value feature and candidate feature subsets, respectively. According to the definition of the mutual information, we will get the mutual information between f_i and f_j by Eq. (1).

$$I(f_i;f_j) = \sum_{f_i \in F} \sum_{f_j \in F j \neq i} p(f_i,f_j) \log \frac{p(f_i,f_j)}{p(f_i)p(f_j)}. \quad (1)$$

Under this context, those features $f_j \in F$ with high predictive power will have larger mutual information $I(f_i;f_j)$. On the contrary, $I(f_i;f_j)$ is zero if f_i and f_j are independent with each other. At this point, f_j has no contribution to the distribution of f_i.

3.2. Finding the most similar cases

In this section, we use the mutual information $I(f_i;f_j)$ as weights to find the most similar cases with the missing values case x_a.

Firstly, in order to discard the feature that is irrelevant or weakly relevant with the selected missing value feature f_i, we first

define a "lower filter": for each selected feature f_i, if $p(I(f_i;f_j) < \varepsilon|n) > \bar{p}, (1 \leqslant i,j \leqslant m)$, we will discard the feature f_i. Here ε is an arbitrary (low) positive threshold and $\bar{p}$ is an arbitrary (high) probability. The feature satisfying the "lower filter" has lower dependency among the candidate feature subsets.

Then, we define the Eq. (2) to find the most similar cases with the missing value case x_a. For convenience, assume that different missing value is in different case. So the we can suppose that k missing value appears in k cases.

$$E_{ab} = \sum_{a=1}^{k} \sum_{b=1,b\neq a}^{n} \sum_{i,j=1}^{m} I(f_i;f_j)(v_{ai} - v_{bj})^2. \tag{2}$$

Here E_{ab} reflects the similarity between the missing value case x_a and other case x_b in the data set. v_{ai} is the ith feature in the case x_a, and it is missing. v_{bj} is the value of the jth feature in the case x_b,x_b is other case except for x_a. The more similar between the two cases, the value of E_{ab} is the more little.

Finally, rank corresponding E_{ab} according to the each missing value v_{ai}. We will get the top σ cases related with each v_{ai} and named as set s_a.$s_{aq}(1 \leqslant q \leqslant \sigma)$ denotes the qth case in the set s_a, where $\sigma = \left[\frac{1}{m}\sqrt{p_j \times E_{ab} \times n}\right]$, p_j is the probability that v_{aj} appears in the sample space, feature f_j is the feature which has maximum mutual information with the feature f_i.

To sum up, in this section, our objective is to get the set s_a, it is the most similar cases with the missing value case x_a than other ones.

3.3. Estimating the value of missing data

At first, we briefly introduce the Newton interpolation polynomial. Then we describe how to apply the polynomial to get the estimation of missing data.

Given a set of n data nodes $(x_1,y_1),\ldots,(x_n,y_n)$, the interpolation polynomial in the Newton form is a linear combination of Newton basis polynomials $N(x) := \sum_{j=1}^{k} a_j n_j(x)$ with the Newton basis polynomials defined as $n_j(x) := \prod_{i=1}^{j}(x - x_i)$ and the coefficients defined as a_j: = $[y_1,\ldots,y_j]$, where $[y_1,\ldots,y_j]$ is the notation for divided differences (http://en.wikipedia.org/wiki/Divided_differences).

Thus, the Newton polynomial can be written as

$$N(x) := [y_1] + [y_1,y_2](x - x_1) + \cdots + [y_1,\ldots,y_k](x - x_1)(x - x_2) \ldots (x - x_{k-1}). \tag{3}$$

$N(x)$ is the estimation of the missing data x.

In this study, we assume that missing data is v_{ai},x_a is the case with v_{ai}, and f_j is the most dependent on the feature f_i. So our target is to estimate the value of v_{ai} using v_{aj} and set s_a.s_a is the most similar cases set with the case x_a,$s_{aq}(1 \leqslant q \leqslant \sigma)$ denotes the qth case in the set s_a, and s_{aqi} denotes the i-th feature in the case s_{aq}.

Corresponding to $(x_1,y_1),\ldots,(x_n,y_n)$, we have σ data nodes $(s_{a1j},s_{a1i}),\ldots,(s_{a\sigma j},s_{a\sigma i})$, according to Eq. (3), we get the Eq. (4).

$$v_{ai} = N(v_{aj}) := [s_{a1i}] + [s_{a1i},s_{a2i}](v_{aj} - s_{a1j}) + \cdots + [s_{a1i},s_{a2i}\ldots,s_{a\sigma i}](v_{aj} - s_{a1j})(v_{aj} - s_{a2j})\ldots(v_{aj} - s_{a\sigma-1j}). \tag{4}$$

The value of $N(v_{aj})$ is the estimation of the missing data v_{ai}. Repeat for each missing data, we will get the estimation of all the missing data.

3.4. Determining the prior feature ordering

As above mentioned, K2 algorithm is very efficient and effective in constructing the Bayesian network, but it need a prior feature ordering (Cooper & Herskovits, 1992; Estevam, Hruschka, Nelson, & Ebecken, 2007), which mainly depends on the subjective experience of doctor and which will make strong effect on the analytical results. In this study, we present a new method to get the prior feature ordering by defining a "higher filter". The specific method is following:

The fundamental target of MCI system is to diagnose MCI. Therefore, we define the functional feature MCI as the root of the Bayesian network, and it has not the parents. That is to say, we set the first feature in the ordering is the functional feature MCI. In order to find the ordering of other features, we set a "higher filter" that is: if $p(I(f_i;f_j) > \varepsilon|n) > 1 - \bar{p}$, we will include and rank these features according to the results of the probability distribution, the definition of notations is same with "lower filter". The magnitude of probability distribution $p(I(f_i;f_j) > \varepsilon|n)$ means the dependent extent among features. If the probability $p(I(f_i;f_j) > \varepsilon|n)$ is same among several features, we will rank them according to the mutual information between the functional feature MCI and the current selected feature f_i. Thus we will get the prior feature ordering. The detailed procedure please sees the Section 5.2.

3.5. Constructing Bayesian network

We adopt the famous K2 algorithm to construct the Bayesian network (Cooper & Herskovits, 1992). Assume f_i has any state in $\{r_1,r_2,\ldots r_m\}$.

In this section, our primary goal is to use Eq. (5) as the score metric to find a Bayesian network B_s that maximize $P(B_s,D)$. Using a simple greedy-search algorithm, we begin the algorithm by assuming a node has no parents, then according to the above gotten prior feature ordering, adds incrementally that parent whose addition most increases the probability of the resulting structure B_s.

$$\max_{B_s}[P(B_s,D)] = \prod_{i=1}^{n} \max_{\pi_i} \left[\prod_{j=1}^{q_i} \frac{(r_i-1)!}{(N_{ij}+r_i-1)!} \prod_{t=1}^{r_i} \alpha_{ijt}!\right]. \tag{5}$$

Here α_{ijt} is the number of cases in the data set D for which f_i = t and $\pi_i = j.N_{ij} = \sum_{t=1}^{r_i}\alpha_{ijt}$. π_i is parent nodes of f_i. Let ϕ_i denote a list of the unique parents of f_i as seen in D. If f_i has no parents, then we define ϕ_i to be the list ϕ, where ϕ represents the empty set of parents. Let q_i = $|\phi_i|$.

By computing Eq. (5) we get the most optimal Bayesian network B_s.

4. Implementation

4.1. The procedure of the MNBN algorithm

The procedure of the MNBN algorithm showed in Fig. 1.

4.2. The procedure of the MCI test

We have recently been exploring a specific MCI diagnostic system applying network analysis, which assists doctor and patients to diagnose the MCI and track the development of the MCI. Furthermore, the system will help patients to understand the MCI based on the clinical data. The first important function of the system is to select intelligently MCI, which is the fundamental target of this study.

The criteria of MCI in this study is provided by Petersen et al. (1999) : (1) memory complaint and corroborated by an informant, (2) normal activities of daily living, (3) normal general cognitive function, (4) objective memory impairment for age, and (5) not demented.

In our data sets, all participants were recruited from the department of neurology of Dalian University affiliated Xinhua hospital (China) after providing informed written consent. We choose the complete data as the experimental data, which includes 45 normal

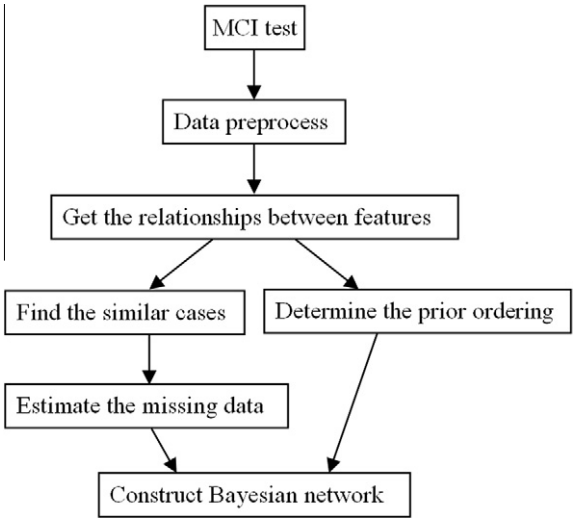

Fig. 1. The MNBN algorithm procedure.

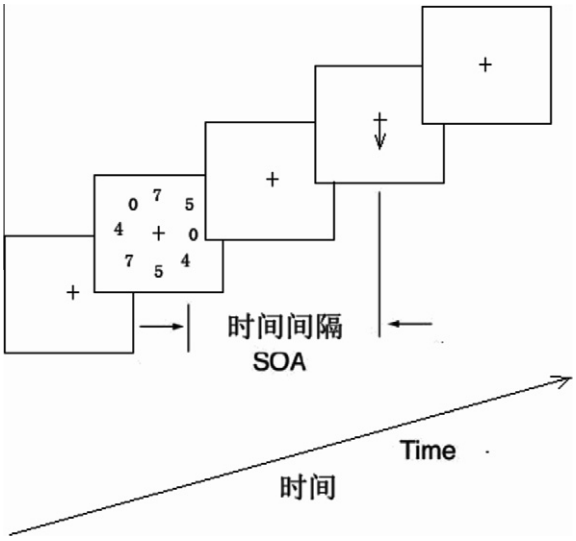

Fig. 2. Display sequence in the test of the STM.

people (mean age ± S.D., 64.78 ± 5.15) and 42 MCI patients (68.52 ± 8.32) were classified by neurological doctor according to the Peterson's criteria. The groups were relatively well-balanced in terms of sex (58% and 55% women in each of the 2 groups, respectively). The procedure of choosing the MCI contains a set of preliminary behavioral tests.

We do not add the brain imaging data into this paper because of the limited space. We focused on the method to infer the decisive factors of MCI. The tests of the system included MMSE (mini-mental state examination) (Folstein, Folstein, & McHugh, 1975), ADL (Activities of Daily Living) (Lawton & Brody, 1969), CDR (Clinical Dementia Rating scale) (Hughes, Berg, Danziger, Coben, & Martin, 1982), ANT (Attentive Networks Test), STM (short-time memory test) and some enquiries of natural information, such as name, No. of ID, age, sex and education degree. None of them was receiving psychoactive medications such as antipsychotic drugs or cerebral vasodilators, nor showed any neurological symptoms or other physical disorders. Moreover, computed tomography (CT) was applied to rule out other organic brain diseases. Considering the purpose of this paper is to describe the method of assistant diagnosis, we do not give the experimental detail. We only give several specific implemental procedures showed as Figs. 2 and 3.

The STM was developed in the environment of the E-Prime, which is commercial experiment software, on an IBM-compatible personal computer. At first, there appears "+"400 ms on the screen. Then appears target stimulation 105 ms. Next, appears an arrow, which pointed randomly to one of the eight numbers. The time internal (target-cue onset asynchrony, SOA) is 11 ms, 32 ms, 74 ms, 516 ms and 1105 ms, respectively. The accuracy of subjects was recorded in every time internal. The experimental procedure is simply showed in the Fig. 2.

ANT was run via E-Prime on an IBM-compatible personal computer. The stimuli consisted of a target and four flankers displayed on a computer screen. The target was a leftward or rightward arrow at the center. This target was flanked on each side by two arrows in the same direction (congruent condition), in the opposite direction (incongruent condition), or by lines without arrowheads (neutral condition). Each target was preceded by some kind cues, such as asterisk, fixation cross and subtraction sign.

The participant was to respond as quickly and accurately as possible based on the direction of the target by pressing the corresponding left or right key on a mouse. The reaction time (RT) and accuracy were recorded. Mainly procedure is showed in Fig. 3.

In the above tests, the goal is to find the influential factors on the MCI. However, some of data can not be obtained in the process of the experiment because the eyesight of participants is weak. So we presented the method to intelligent infer with missing data.

4.3. Overview of the MNBN algorithm

Input: A data set D contains n discrete cases $x_a(1 \leqslant a \leqslant n)$, each case x_a has m features $f_i(1 \leqslant i \leqslant m)$, denoted as v_{ai} and some v_{ai} is missing, the number is k.

Output: Bayesian network B_s.

Step 1: Data preprocess We chose the following 9 features: age, sex, education degree, CDR score, MMSE score, ADL score, ANT, STM and MCI functional feature.

For the data of the STM, we adopt the mean of the accuracy in the 0 ms, 116 ms, 137 ms, 179 ms, 621 ms, 1210 ms as the results of the STM. For the data of the ANT, we observed the accuracy of most subjects is nearly 100%, so we adopt the mean of the react time as the results of the ANT. Other data was discretized respectively.

Step 2: Get the relationships $I(f_i;f_j)$ between the feature f_i and f_j:

According to Eq. (1), Repeat for each feature $f_i,f_j \in F$ calculating the mutual information:

$$I(f_i;f_j) = \sum_{f_i \in F} \sum_{f_j \in F, j \neq i} p(f_i,f_j) \log \frac{p(f_i,f_j)}{p(f_i)p(f_j)}.$$

Step 3: Find the similar cases set s_a:

1. Apply the "lower filter" $p(I(f_i;f_j) < \varepsilon | n) > \bar{p}$ to discard those irrelevant or weakly relevant features.
2. Repeat for each case and each feature computing

$$E_{ab} = \sum_{a=1}^{k} \sum_{b=1,b\neq a}^{n} \sum_{i,j=1}^{m} I(f_i;f_j)(v_{ai} - v_{bj})^2.$$

3. Rank these E_{ab}, find the top σ cases and named as set s_a corresponding each missing data v_{ai}, $\sigma = [\frac{1}{m}\sqrt{p_j \times E_{ab} \times N}]$.

Step 4: Estimate the value of missing data v_{ai} by Computing:

$$v_{ai} = N(v_{aj}) := [s_{a1i}] + [s_{a1i}, s_{a2i}](v_{aj} - s_{a1j}) + \cdots + [s_{a1i}, s_{a2i} \ldots, s_{a\sigma i}](v_{aj} - s_{a1j})(v_{aj} - s_{a2j}) \ldots (v_{aj} - s_{a\sigma-1j}).$$

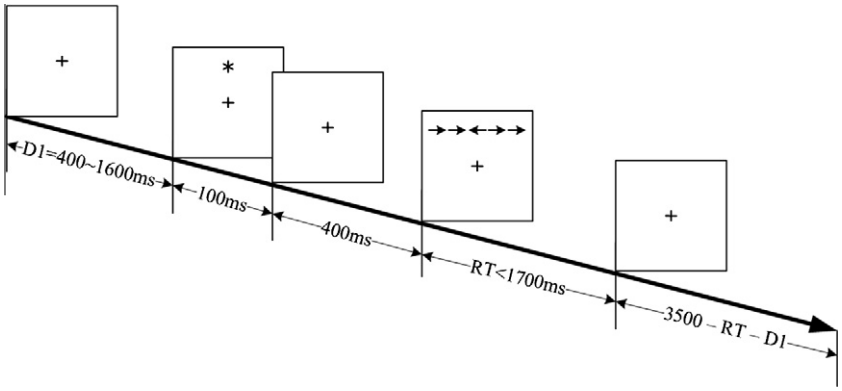

Fig. 3. Display sequence in the test of the ANT.

Repeat for each missing data v_{ai}, we will get the estimation of all the missing data.

Step 5: Determine the prior feature ordering P_{red}.

1. Set MCI as the first feature node in the predicted feature ordering P_{red}.
2. Apply the "higher filter" $p(I(f_i;f_j) > \varepsilon|n) > 1-\bar{p}$ to include and rank the features.

Step 6: Construct the Bayesian network B_s (the set π_i)

1. Repeat for each case, set initialized parameter $\pi_i = \phi$;

$$P_{old} = K_2(f_i, \pi_i) = \prod_{j=1}^{q_i} \frac{(r_i-1)!}{(N_{ij}+r_i-1)!} \prod_{t=1}^{r_i} \alpha_{ijt}!.Flag = true.$$

2. Iterate for each $flag = true$ and $|\pi_i| < u$, update $z = P_{red}(f_i) - \pi_i$ that maximizes $K_2(f_i, \pi_i \cup \{z\})$ and $P_{new} = K_2(f_i, \pi_i \cup \{z\})$. Here $P_{red}(f_i)$ is the prior feature ordering and u is the permitted maximal number of parent nodes, in general, u=3.
3. If $(P_{new} > P_{old})$ then $P_{old} = P_{new}$; and $\pi_i = \pi_i \cup \{z\}$.

5. Experimental results

5.1. Experiments on the MCI

The algorithm has been implemented in MATLAB. All of the experiments are conducted on the IBM personal computer with 2.0 GHz processor and 3 GB memory running Windows XP operating system.

In order to compare the robustness of the response models, we adopt a 5-fold cross-validation approach for performance estimation. A data set is randomly partitioned into 5 mutual exclusive and exhaustive folds. Each time, a different fold is chosen as test set and other four folds are combined together as the training set. Response models are learned from the training set and evaluated on the corresponding test set.

For convenience in later discuss, in the MCI experiment, we define 9 features from 1 to 9, which are age, sex, education degree, MMSE, ADL, CDR, STM, ANT and MCI, respectively. The Table 1 describes the mutual information among 9 features in the training set. We can see that the 3rd, 7th and 8th feature are high dependent on the MCI functional feature, so we should give more weight in finding the similar cases. Meanwhile, we can see that the 2nd feature satisfying the "lower filter", therefore, in the process of estimating the missing data and constructing the Bayesian network, we discard the feature, which will largely decrease the complex of the computation, especially in the large scale data sets.

The specific results are given in Table 1.

Next, we estimate missing data in the MCI test set. According to the step 2, step 3 and step 4 of Section 4, we get the estimation of the missing data. Specific results are showed in the Table 3.

In order to improve the efficiency, in finding the similar cases, we do not find the most similar cases in the global data set. In general, after we find $\frac{n}{k}$ missing values, we will get the minimal error between the missing data cases and other cases, then in the following experiment, we find those cases less than or equal to minimal error as the similar cases, which will save much search time and get the better results.

According to the step 5 and step 6 of the Section 4, we construct the Bayesian network of MCI showed in Fig. 4. The network shows that the feature sex is weak dependent on all other features and therefore not shown in Fig. 4. The Bayesian network in Fig. 4 represents multivariate nonlinear associations among 8 features and the functional feature MCI. Given the results for a set of tests, we can compute the posterior distribution for the clinical feature MCI from the Fig. 4. That is, we can predict the state of feature MCI based on the results of the tests. For example, MCI is directly dependent on ANT, STM and education degree. In this way, for each subjects we can compute the probability $p = P(MCI|ANT, STM, educationdegree)$, based on the results of the ANT, STM test and education degree. If $p < 0.2$, we can predict the state of feature MCI as 0 (Normal). If $0.2 \leqslant p \leqslant 0.5$, we can predict the state of MCI as

Table 1
The mutual information among features in the MCI.

	f_1	f_2	f_3	f_4	f_5	f_6	f_7	f_8	f_9
f_1	8.3246	0.6352	0.9845	2.3467	5.6432	3.1675	2.4563	5.6749	3.2135
f_2	0.6352	6.4375	1.2364	0.3495	1.2436	0.9358	2.1038	1.4573	0.7366
f_3	0.9845	1.2364	34.3465	5.3429	2.4685	3.4256	3.6235	4.5323	6.4367
f_4	2.3467	0.3495	5.3429	12.3483	3.2675	3.2643	3.2498	4.3847	2.3646
f_5	5.6432	1.2436	2.4685	3.2675	15.4832	3.2476	1.6473	4.3728	3.3235
f_6	3.1675	0.9358	3.4256	3.2643	3.2476	25.3594	2.1324	5.4635	3.1242
f_7	2.4563	2.1038	3.6235	3.2498	1.6473	2.1324	18.2394	3.2421	5.4382
f_8	5.6749	1.4573	4.5323	4.3847	4.3728	5.4635	3.2421	42.3421	9.3438
f_9	3.2135	0.7366	6.4367	6.3646	3.3235	3.1242	5.4382	9.3438	62.4386

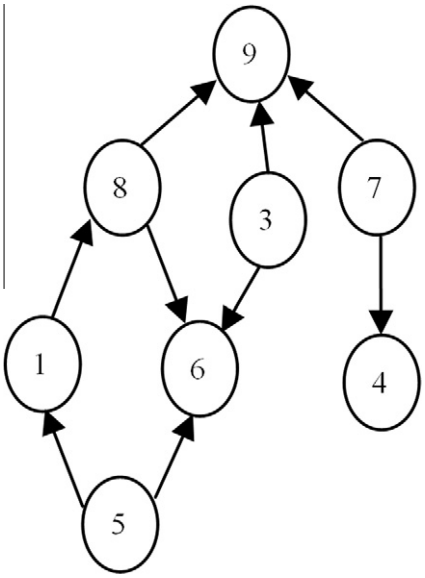

Fig. 4. The Baysian network of MCI.

1(MCI). If $p > 0.5$, we can predict the state of MCI as 2(MCI to dementia).

The predictive accuracy of using ANT, STM and education degree jointly to predict MCI is 0.82(sensitivity 0.79 and specificity 0.84). If we use these three features independently, the accuracy of ANT is 0.72(sensitivity 0.69 and specificity 0.76), that of STM is 0.68 (sensitivity 0.74 and specificity 0.62, that of education degree is 0.56(sensitivity 0.62 and specificity 0.51). This finding suggests the importance of jointly considering the states of the ANT, STM and education degree in diagnosing MCI.

On the other side, we can conclude that if we want to do a primary diagnosis quickly, we may only test the ANT,STM and combine with the information of education degree, we will get the diagnosis with highly accuracy in short time. We might apply this idea into the hospital to check the body for the general people.

5.2. Experiments on the benchmark data sets

In order to further validate the performance of the MNBN algorithm, we test the algorithm on the six standard data sets. We first take the Pima Indians Diabetes data set (PID, one of the data set in the UCI repository (Newman et al., 1998)) as example to further introduce the procedure of the MNBN algorithm. Then, we verify the performance of the MNBN algorithm by testing on the MCI and six data sets from UCI repository.

The PID data set contains 768 cases and 9 features. The meaning of each feature is: 1. Number of times pregnant; 2. Plasma glucose concentration a 2 h in an oral glucose tolerance test; 3. Diastolic blood pressure (mm Hg); 4. Triceps skin fold thickness (mm); 5. 2-h serum insulin (mu U/ml); 6. Body mass index (weight in kg/(height in m)^2); 7.Diabetes pedigree function; 8. Age (years); 9. Diabetes (0 or 1). For convenience in later discuss, we named these 9 features from 1 to 9, respectively.

According to Eq. (1), we first get the results of the mutual information in the Table 2.

Then, we generate randomly the missing data in the data set. According the step 2, step 3 and step 4 in Section 4 we obtain

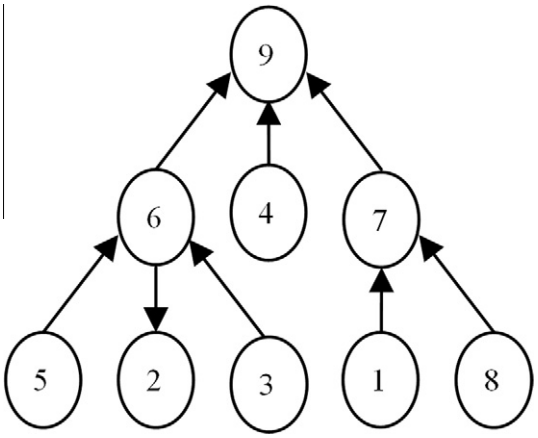

Fig. 5. The Baysian network of the PID.

Table 2
The mutual information among all the features in the PID.

	f_1	f_2	f_3	f_4	f_5	f_6	f_7	f_8	f_9
f_1	172.4335	1.7566	1.6407	1.7662	1.1030	1.9076	2.1098	4.3449	4.5956
f_2	1.7566	36.6211	2.1604	2.1848	2.3484	3.4966	4.1819	2.2351	1.3867
f_3	1.6407	2.1604	101.2492	2.9478	2.3315	2.6304	2.8798	2.0566	0.7475
f_4	1.7662	2.1848	2.9478	124.4140	49.6010	2.5099	2.9127	1.5300	0.1954
f_5	1.1030	2.3484	2.3315	49.6010	140.3182	2.6317	3.1042	1.2800	0.3969
f_6	1.9076	3.4966	2.6304	2.5099	2.6317	23.8376	4.7661	2.5470	0.8147
f_7	2.1098	4.1819	2.8798	2.9127	3.1042	4.7661	11.4355	3.0732	0.6703
f_8	4.3449	2.2351	2.0566	1.5300	1.2800	2.5470	3.0732	93.4862	4.1898
f_9	4.5956	1.3867	0.7475	0.1954	0.3969	0.8147	0.6703	4.1898	238.1666

Table 3
The performance of the MNBN algorithm in different data sets.

	Iris	Wine	PID	Cloud	Statlog	MGT	MCI
No. of features	5	14	9	11	37	11	9
No. of complete cases	143	163	729	961	3750	18,351	78
No. of missing data	10	20	50	100	1000	1000	11
No. of incomplete cases	7	15	39	63	685	669	9
Missing percent (%)	1.33	0.80	0.72	0.88	0.61	0.48	1.39
MSE	0.0158	0.0667	0.2280	0.0319	3.1713	0.5919	0.0173
AET	0.0131	0.3097	2.2967	3.1158	377.1224	687.3553	0.0136

Table 4
The performance comparison among LibB, EBN and MNBM.

Missing percent (%)	No. of missing values	No. of incomplete cases	Method	ASD	AET
0.1	370	322	LibB	27.8 ± 15.3	379.3 ± 103.8
			EBN	6.1 ± 5.8	351.1 ± 108.6
			MNBN	5.6 ± 4.7	366.3 ± 29.3
1	3700	2826	LibB	27.4 ± 15.8	821.5 ± 223.9
			EBN	7.2 ± 4.4	694.3 ± 159.1
			MNBN	6.9 ± 3.6	839.5 ± 34.2
5	18.500	7639	LibB	29.7 ± 10.0	3012.2 ± 789.6
			EBN	8.6 ± 4.3	1553.8 ± 252.7
			MNBN	8.2 ± 4.1	2178.6 ± 39.8

the estimation of the missing data. Specific results are showed in the Table 3.

Next, we will describe detailed procedure that mutual information determines the prior feature ordering P_{red} according to the step 5. The detailed procedure is followed:

1. Set MCI as the first feature node in the predicted feature ordering P_{red}.
2. Apply the "lower filter" $p(I < \varepsilon|n) > \bar{p}$. We can not discard any feature because no any feature satisfying the "lower filter". $\varepsilon = 2.0, \bar{p} = 0.75$.
3. Apply the "higher filter" $p(I > \varepsilon|n) > 1 - \bar{p}$ we will get all the features which satisfy the "higher filter". We rank these features. The ordering is $f_9,f_7,f_2,f_3,f_6,f_8,f_4,f_5,f_1$. However, among them the probability distribution of f_2,f_3,f_6,f_8 and f_4,f_5 is same, respectively. So we rank them according to the mutual information between the current feature and the functional feature f_9. Therefore, we will get the feature ordering P_{red} is $f_9,f_7,f_8,f_2,f_6,f_3,f_5,f_4,f_1$.

Finally, according to the step 6, apply the feature ordering P_{red} to the K2 algorithm, we construct the diagnostic network of the PID showed in Fig. 5.

From the Fig. 5, we can conclude that strong association among the Pima Indian Diabetes and body mass index (weight/height) and triceps skin fold thickness. The positively correlated with body mass index (weight/height) and blood pressure, diabetes pedigree function and ages, respectively, which is consistent with other reports (Knowler, Bennett, Hamman, & Miller, 1978; de Courten, Pettitt, & Knowler, 1996).

In order to further evaluate the effectiveness of the MNBN algorithm, MCI and six data sets from UCI repository (Newman et al., 1998) were used. These data sets were Iris, Wine, Pima Indians Diabetes (PID), Cloud, Landsat Satellite (Statlog) and Magic Gamma Telescope (MGT).

We evaluate the performance of the algorithms using the following parameters:

No. of missing data: equal to size of all cases × No. of features × missing percentage.
MSE: mean square error is the sum of (real value-estimated value)^2, then divides k.
AET: the average execution time of each data set in seconds.

From the Table 3, we can conclude that most of the results in MSE can be accepted, only minimum MSE is a little high, such as Statlog. We analyze the reason that the range of data is larger and difference between the data is larger. On the other hand, some estimations of missing data are very precise, such as the MSE can arrive at 0.0027 in the 5th feature of the MGT data set even with 1000 missing value. In the computation of the MSE, we adopt the mean of all the MSE.

5.3. Comparisons among the MNBN and other algorithms

In order to compare the performance of MNBN with other algorithms, including EBN (Wong & Guo, 2008) and LibB,[1] we tested the different methods on the well-known benchmark network the Alarm data set (Beinlinch, Suermondt, Chavez, & Cooper, 1989; Cooper & Herskovits, 1992) with different missing percentage. LibB can learn Bayesian networks from data in the presence of missing data, which implements the SEM algorithm (Friedman, 1998) introduced in Section 2.

Firstly, we randomly sample the original data set from the alarm with no missing data. The No. of case is 10,000, No. of feature is 37. Then, the incomplete data set used in our experiments are generated from the corresponding original data set with missing data introduced randomly. Because the EBN and LibB are two stochastic algorithms, we execute them for 20 times on each data set to get their average performance. The Table 4 is the average and standard deviation of 20 trails.

ASD: The average structural difference, i.e., number of edges added, reversed and omitted, between the final solution and the original network structure.

From the Table 4, we can draw the following conclusions:

(1) The precision of the MNBN algorithm is better than SEM and EBN algorithm, although time efficiency is worse than other algorithms in some times.
(2) From the results of the standard deviation, The MNBN algorithm is more stable than other algorithms. The percentage of missing data makes little effect to the MNBN algorithm. We analyze the reason that may be the MNBN algorithm has adaptive ability through finding the similar cases.

6. Conclusions

In this paper, we use a set of behavioral experimental data constructing the MCI network model. We find ANT, STM and education degree are the mainly influencing factors of MCI, and get the nonlinear association among these influencing factors.

The MCI system has been tested and applied in Xinhua hospital of Dalian, and we have confidence that it has potential to apply in society and ordinary family. Because each person can detect the possibility of MCI risks at home, which will greatly improve the discovery rate of MCI. If the subject is detected the high MCI risk, the system may advice the subject to go to hospital and do the further examination under the guidance of doctor. So our system might be a primary free diagnosis to MCI.

In the process of constructing MCI network, we propose the MNBN algorithm, which first uses the mutual information between features to find the similar cases with the missing data, adopt Newton interpolation to estimate the missing data. Next, we utilize

[1] LibB is available at http://compbio.cs.huji.ac.il/LibB/.

again the mutual information and define "higher filter" to get the suitable feature ordering. Finally, we apply the feature ordering to construct the Bayesian network. The experimental results indicate that MNBN algorithm achieves better results than other methods in most conditions.

In the future, we plan to further improve the efficiency of the algorithm by following several aspects:

(1) In the process of finding the similar cases, considering the complexity, we currently only compute the mutual information between two features and ignore the effects of the selected feature subset. In some conditions, if necessary, we could utilize the multivariate mutual information to further improve the accuracy. By using multivariate mutual information, we will get the sets of dependent relationships among a set of features. But it will increase the complexity of the algorithm. Therefore, how to utilize some optimal methods to increase the efficiency of the algorithm is a challenge and require further investigation intensely.
(2) Using Newton interpolation to estimate the missing data demands the feature value different from each other. In order to decrease the complexity, we only compute the mean of corresponding feature values, which will make effect on the precision of the algorithm. In future, we will develop more optimal algorithm to estimate the missing data.
(3) Although we gain the better performance on the above data sets, more studies will be required on how to tune the parameters, such as σ, "higher filter" and "lower filter", which is a problem worth investigating in future research.

Acknowledgement

This work was supported in part by the National Natural Science Foundation of China (No. 60971096). The authors thank the members of the department of neurology of Dalian University affiliated Xinhua hospital for their invaluable help with this study.

References

Austin, P., & Escobar, M. (2005). Bayesian modeling of missing data in clinical research. *Computational Statistics & Data Analysis, 49*(3), 821–836.

Beinlinch, I., Suermondt, H., Chavez, R., & Cooper, G. (1989). The ALARM monitoring system: A case study with two probabilistic inference techniques for belief networks. *Proceedings of Second European Conference Artificial Intelligence in Medicine*, 247–256.

Celsis, P. (2000). Age-related cognitive decline, mild cognitive impairment or preclinical Alzheimer's disease? *Annals of Medicine, 32*(1), 6–14.

Chen, R., & Herskovits, E. H. (2006). Network analysis of mild cognitive impairment. *NeuroImage, 29*, 1252–1259.

Cooper, G. F., & Herskovits, E. H. (1992). A Bayesian method for the induction of probabilistic networks from data. *Machine Learning, 9*, 309–347.

de Courten, M. P., Pettitt, D. J., & Knowler, W. C. (1996). Hypertension in Pima Indians: Prevalence and predictors. *Public Health Reports, 111*(2), 40–43.

Dempster, A. P., Laird, N. M., & Rubin, D. B. (1977). Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society. Series B (Methodological), 39*(1), 1–38.

Estevam, R., Hruschka Jr., Nelson, F. F., & Ebecken (2007). Towards efficient variables ordering for Bayesian networks classifier. *Data & Knowledge Engineering, 63*(2), 258–269.

Folstein, M. F., Folstein, S. E., & McHugh, P. R. (1975). "Mini-mental state". A practical method for grading the cognitive state of patients for the clinician. *Journal of Psychiatric Research, 12*(3), 189–198.

Friedman, N. (1997). *belief networks in the presence of missing values and hidden variables*. *Proceedings of the 14th International Conference on Machine Learning*. Morgan Kaufmann (pp. 125–133).

Friedman, N. (1998). The Bayesian structural EM algorithm. In *Proceedings of the 14th conference on uncertainty in artificial intelligence. San Francisco* (pp. 129–138).

Garcia-Laencina, Pedro J., Sancho-Gomez, Jose-Luis, Figueiras-Vidal, Anibal R., & Verleysen, Michel. (2009). K nearest neighbours with mutual information for simultaneous classification and missing data imputation. *Neurocomputing, 72*(7–9), 1483–1493.

Heckerman, D., Geiger, D., & Chickering, M. (1994). Learning Bayesian networks: The combination of knowledge and statistical data. In *Proceedings of the 10th conference on uncertainty in artificial intelligence*.

Hughes, C. P., Berg, L., Danziger, W. L., Coben, L. A., & Martin, R. L. (1982). A new clinical scale for the staging of dementia. *The British Journal of Psychiatry, 140*, 566–572.

Knowler, William C., Bennett, Peter H., Hamman, Richard F., & Miller, Max (1978). Diabetes incidence and prevalence in pima Indians: A 19-fold greater incidence than in Rochester, Minnesota. *American Journal of Epidemiology, 108*(6), 497–505.

Lawton, M. P., & Brody, E. M. (1969). Assessment of older people: Self-maintaining and instrumental activities of daily living. *Gerontologist, 9*, 179–186.

Liang, F., & Zhang, J. (2009). Learning Bayesian networks for discrete data. *Computational Statistics and Data Analysis, 53*, 865–876.

Lin, J. H., & Haug, Peter J. (2008). Exploiting missing clinical data in Bayesian network modeling for predicting medical problems. *Journal of Biomedical Informatics, 41*, 1–14.

Liu, H., Sun, J., Liu, L., & Zhang, H. (2009). Feature selection with dynamic mutual information. *Pattern Recognition, 42*(7), 1130–1339.

Marcus Hutter & Zaffalon, M. (2005). Distribution of mutual information from complete and incomplete data. *Computational Statistics and Data Analysis, 48*, 633–657.

Morris, J. C., Storandt, M., Miller, J. P., McKeel, D. W., Price, J. L., Rubin, E. H., et al. (2001). Mild cognitive impairment represents early-stage Alzheimer disease. *Arch. Neurol, 58*(3), 397–405.

Myers, J. W., Laskey, K. B., & DeJong, K. A. (1999). *Learning Bayesian networks from incomplete data using evolutionary algorithms*. *Proceedings of the Fourth Annual Conference on Genetic and Evolutionary Computation Conference*. Orlando, Florida: Morgan Kauffman (pp. 458–465).

Newman, D.J., Hettich, S., Blake, C.L., & Merz. C.J. (1998). UCI repository of machine learning databases. Department of Information and Computer Science, University of California, Irvine. <http://www.ics.uci.edu/~mlearn/MLRepository.html>.

Pena, J. M., Lozano, J. A., & Larranaga, P. (2000). An improved Bayesian structural EM algorithm for learning Bayesian networks for clustering. *Pattern Recognition Letters, 21*, 779–786.

Pena, J. M., Lozano, J. A., & Larranaga, P. (2002). Learning recursive Bayesian multinets for data clustering by means of constructive induction. *Machine Learning, 47*, 63–89.

Petersen, R. C., Smith, G. E., Waring, S. C., Ivnik, R. J., Tangalos, E. G., & Kokmen, E. (1999). Mild Cognitive Impairment: Clinical Characterization and Outcome. *Arch Neurol, 56*, 303–308.

Petersen, R. C., Stevens, J. C., Ganguli, M., Tangalos, E. G., Cummings, J. L., & DeKosky, S. T. (2001). Parameter: Early detection of dementia: Mild cognitive impairment. *Report of the quality standards subcommittee of the American Academy of Neurology*, 1133–1142.

Wong, M. L., & Guo, Y. Y. (2008). Learning Bayesian networks from incomplete databases using a novel evolutionary algorithm. *Decision Support Systems, 45*, 368–383.

Wong, M. L., & Leung, K. S. (2004). An efficient data mining method for learning Bayesian networks using an evolutionary algorithm-based hybrid approach. *IEEE Transactions on Evolutionary Computation, 8*(4), 378–404.