International Journal of Innovative
Computing, Information and Control
Volume **9**, Number **11**, November **2013**

ICIC International ©2013 ISSN 1349-4198
pp. 4545–4554

INTELLIGENT DIAGNOSTIC SYSTEM FOR CEREBROVASCULAR DISEASES BASED ON A BAYESIAN NETWORK WITH INFORMATION GAIN

Yan Sun[1,*], Ying Bai[2], Shuxue Ding[3], Yi-Yuan Tang[4,5]
Yifen Cui[4] and Yan Wang[4]

[1]School of Psychology
Liaoning Normal University
No. 850, Huanghe Road, Shahekou District, Dalian 116029, P. R. China
[*]Corresponding author: sunyan@lnnu.edu.cn

[2]Department of Neurology
Dalian University Affiliated Xinhua Hospital
No. 156, Wansui Street, Shahekou District, Dalian 116021, P. R. China

[3]School of Computer Science and Engineering
Aizu University
Aizu-Wakamatsu City, Fukushima 965-8580, Japan

[4]Institute of Neuroinformatics
Dalian University of Technology
No. 2, Linggong Road, Ganjingzi District, Dalian 116024, P. R. China

[5]Texas Tech Neuroimaging Institute and Department of Psychology
Texas Tech University
Lubbock, TX 79409, USA

Received November 2012; revised March 2013

Abstract. *In this paper, we present an intelligent system for analyzing the probabilistic dependencies that valuate the relationships of risk factors of cerebrovascular diseases (CVDs). We demonstrate the process used by the system to diagnose CVDs. To construct the system, we select age, gender, hypertension, diabetes mellitus, coronary heart disease, and hyperlipemia as risk factors of CVDs, which are based on the advice of experienced CVD doctors. The associations of CVDs with these risk factors are analyzed. To diagnose CVDs based on these risk factors objectively, we propose a novel system model based on a Bayesian network (BN) and information gain. By training the model using standard datasets, we obtain a diagnosis system that can automatically generate a diagnosis result when a group of data incorporating the risk factors is inputted. Finally, we test and evaluate the system using standard datasets and compare the results with those of support vector machine analysis. We also present the evaluation results from three experienced CVD doctors, who confirm that the diagnosis results of the system are beneficial to the realistic diagnosis and prediction of CVDs.*
Keywords: Bayesian network (BN), Cerebrovascular diseases (CVDs), Information gain, Risk factor

1. **Introduction.** Cerebrovascular diseases (CVDs) are major causes of morbidity and mortality worldwide [1]. CVDs are reportedly some of the leading factors of death caused by human disease. The incidence rate ranges from 1‰ to 3‰ and 60% to 70% are disabled among survivors worldwide. In the Asian population, the incidence rate of CVDs ranges from 3% to 5%. CVDs pose a serious threat to human health [2], in addition to being associated with high medical expenses and long term health care burden. Therefore, the

diagnosis of CVDs and the determination of relationships among the risk factors of CVDs are important.

Many researchers are committed to studying the risk factors of CVDs. Basic statistical and epidemiological methods have revealed that hypertension, diabetes mellitus, coronary heart disease, and hyperlipemia are the primary risk factors of CVDs [2]. Hypertension is the highest [2] and hyperlipemia is the second highest risk factor. Some studies have also focused on the relationships among these risk factors. For example, diabetes mellitus has been found to be associated with a marked increase in coronary heart disease [3,4].

Some researchers have recently recognized the importance of CVD prediction and determination of relationships among CVDs and related risk factors. Yeh [5] proposed a predictive system for analyzing the eight important risk factors of CVDs, from which he extracted 16 diagnosis classification rules. These rules are based on classification algorithms and neural networks but are very complicated to apply in general cases even with modifications. To solve this problem, we present a Bayesian network (BN)-based method to reveal the complex, nonlinear multivariate associations among CVDs and related risk factors. One of the main reasons for using BN to diagnose and predict CVDs is that the relationship among the risk factors, whether significant or subtle, can be clearly observed and quantified. These relationships are not fixed as rules based on simple comparisons. The CVD system can also make probabilistic inferences, and full joint distribution is not required in the process. With these advantages, BNs have been successfully applied in many domains in recent years [6,7].

In this paper, information gain technology is introduced to construct an optimized BN. Specifically, we construct a BN system that can diagnose and predict CVDs, as well as provide a dependable analysis of CVDs and related risk factors. In developing the network-based model, we use the K2 algorithm [8] to construct the network. Considering that the result of this algorithm can be influenced by the ordering of the inputted attributes, the optimization of the ordering has also been studied [7]. To optimize the method, we propose an efficient method adopting information gain technology. This method uses a "queue device" for obtaining the prior attribute ordering, which increases the objectivity and effectiveness of the model.

The remainder of this paper is organized as follows. The proposed model and its details are explained in Section 2. The experimental results are provided in Section 3 to show the usefulness and validity of the proposed method. A discussion on the proposed method is found in Section 4. Section 5 summarizes the paper.

2. **Methods.**

2.1. **Dataset.**

The dataset consists of 825 patients from the Department of Neurology of Dalian University, which is affiliated with the Xinhua Hospital (China). The criteria for CVD diagnosis published by the World Health Organization in 1999 are adopted. Based on the suggestions of clinical experts, the risk factors of CVDs mainly include age, gender, hypertension, diabetes mellitus, coronary heart disease, and hyperlipemia. In this paper, we use BN to model the interactions among these six risk factors and cerebral infarction (CVD). In formulation, we denote the six risk factors and CVD as X_1, X_2, X_3, X_4, X_5, X_6, X_7. We discretize the attributes because some of them have continuous values and the BN requires discrete states. Based on the data distribution, we discretize the attributes X_1, X_2, X_3, X_4, X_5, X_6, X_7 into grades 5, 2, 6, 6, 6, 2 and 2, respectively. In this study, the BN is a probabilistic graphical model whose nodes represent six risk factors and CVD. The edges indicate direct conditional dependencies between the connected nodes.

2.2. **Determining the initial attribute ordering.** We construct the BN using the popular and efficient K2 algorithm [8]. However, a prior sequence of the nodes is required to use this algorithm [8]. For example, the node of any descendant cannot appear earlier than the parent nodes in the node ordering [6]. In general, the prior node ordering is specified based on professional knowledge or subjective experience, which significantly affects the results of the BN model. Some researchers have constructed a maximum-weight spanning tree to obtain the prior ordering [6,9]. In this paper, we use the results of the information gain and a "queue device" to optimize the prior node ordering. For the convenience of description, the nodes of BN are named as attributes, which represent CVDs and related risk factors.

2.2.1. *Information gain.* The information gain $IG(X_j; X_i)$ of a given attribute X_i with respect to another attribute X_j is the reduction in uncertainty of the value of X_j when we know the value of X_i (where $i, j = 1, \ldots, n$), and n is the number of attributes in the dataset. The information gain describes the quantity of information that the attribute brings relative to that of the other attributes in the system. The uncertainty of the value of X_j is measured based on its entropy $H(X_j)$ if it is independent on others. The uncertainty of the value of X_j when we know the value of X_i is measured based on the conditional entropy of X_j given X_i, $H(X_j | X_i)$. This information gain can be formulated as $IG(X_j; X_i) = H(X_j) - H(X_j | X_i) = H(X_i) + H(X_j) - H(X_i, X_j)$.

Given that $k = 1, \ldots, m$ and m is the number of cases in the dataset, if X_j and X_i are attributes with values from $\{X_{j1}, \ldots, X_{jk}\}$ and $\{X_{i1}, \ldots, X_{ik}\}$, respectively, the entropy of X_j is obtained as $H(X_j) = -\sum_{k=1}^{m} P(X_j = x_{jk}) \log_2(P(X_j = x_{jk}))$. The conditional entropy of X_j given X_i is obtained as $H(X_j | X_i) = -\sum_{k=1}^{m} P(X_i = x_{ik}) H(X_j | X_i = x_{ik})$.

Therefore, the information gain can be obtained as follows:

$$IG(X_j; X_i) = -\sum_{k=1}^{m} P(X_j = x_{jk}) \log_2(P(X_j = x_{jk})) + \sum_{k=1}^{m} P(X_i = x_{ik}) H(X_j | X_i = x_{ik}).$$

2.2.2. *Queue device.* To determine the ordering of the attributes, we develop a "queue device" that sorts the sequence of the attributes according to $P(IG(X_i; X_j) > \varepsilon)$. Here, ε is arbitrary and usually a high positive threshold. P is a probability distribution of the information gain between two attributes. The attributes in front of the "queue device" are those that bring more information to the system. On the other hand, the attributes at the back of the "queue device" are those that bring relatively less information to the system. If the probability distribution is similar among several attributes, we can sort the attribute based on another value of ε. Table 1 shows an example of the information gains among attributes of a specific data. This example illustrates the attribute ordering by the queue device.

Based on the definition of the "queue device", we set $\varepsilon = 8$. Thus, the content of the "queue device" is as follows: $P(IG(X_1, :) > \varepsilon) = 1/7$, $P(IG(X_2, :) > \varepsilon) = 2/7$, $P(IG(X_3, :) > \varepsilon) = 3/7$, $P(IG(X_4, :) > \varepsilon) = 3/7$, $P(IG(X_5, :) > \varepsilon) = 3/7$, $P(IG(X_6, :) > \varepsilon) = 3/7$, and $P(IG(X_7, :) > \varepsilon) = 4/7$.

After sorting these probabilities, the prior sequence of our system is as follows: $\{X_1, X_2\}$ is at the front of the sequence, $\{X_3, X_4, X_5, X_6\}$ is at the middle, and X_7 is at the back. Considering that the probabilities of some of the attributes are similar in the "queue device", we randomly select $\varepsilon = 6$. The content of the "queue device" is as follows: $P(IG(X_1, :) > \varepsilon) = 1/7$, $P(IG(X_2, :) > \varepsilon) = 3/7$, $P(IG(X_3, :) > \varepsilon) = 3/7$,

TABLE 1. Information gain among attributes

	X_1	X_2	X_3	X_4	X_5	X_6	X_7
X_1	0	3.8599	3.1351	12.3107	4.2470	4.4753	4.2612
X_2	9.2129	0	10.6102	7.8818	4.2606	2.2174	5.6047
X_3	9.6801	8.7032	0	13.1538	3.8201	3.2413	4.1124
X_4	9.2527	9.1091	12.9201	0	4.4217	2.2092	7.0881
X_5	8.8923	8.1004	9.9915	6.9542	0	1.5519	5.6332
X_6	8.6944	8.3198	10.3167	7.1590	4.1023	0	4.6129
X_7	8.8824	8.9512	10.0161	8.6566	5.3500	2.0129	0

$P\left(IG\left(X_4,:\right)>\varepsilon\right)=4/7$, $P\left(IG\left(X_5,:\right)>\varepsilon\right)=4/7$, $P\left(IG\left(X_6,:\right)>\varepsilon\right)=4/7$, and $P(IG\left(X_7,:\right)>\varepsilon)=4/7$.

Thus, the prior sequence of our system is as follows: $\{X_1, X_2, X_3\}$ is at the front of the sequence, and $\{X_4, X_5, X_6, X_7\}$ is at the back. Based on the contents of "queue device", we obtain the prior sequence as follows: $\{X_1, X_2\}$ is at the front of the sequence, $\{X_3, X_4\}$ is at the middle, and $\{X_5, X_6, X_7\}$ is at the back. Certainly, the order of attributes in $\{X_1, X_2\}$ can be a random permutation of these attributes similar to $\{X_3, X_4\}$ and $\{X_5, X_6, X_7\}$.

2.3. **Constructing BN.** In this paper, n risk factors of the problem domain [where X_i $(1 \le i \le n)$] are represented as attributes (nodes) of BN. Each attribute X_i is assumed as any state $\{r_1, r_2, \ldots, r_n\}$. A strong correlation between two attributes is represented as an edge connecting these attributes, which is based on the minimum description length scoring criterion [9]. The joint probability distribution can be computed using Equation (1):

$$P\left(A,B,C\right)=P\left(C|A,B\right)\times P\left(A\right)\times P\left(B\right) \tag{1}$$

To obtain a BN from real application with dataset D, we need to define a scoring metric to describe the fitness between the selected BN model and observed dataset D using Equation (2):

$$\max_{B_s}\left[P(B_s,D)\right]=\prod_{i=1}^{n}\max_{\pi_i}\left[\prod_{j=1}^{q_i}\frac{(r_i-1)!}{(N_{ij}+r_i-1)!}\prod_{t=1}^{r_i}\alpha_{ijt}!\right], \tag{2}$$

which is adopted in literature [6,8]. Here, B_s is the structure of BN, α_{ijt} is the number of cases in the dataset for which $X_i = t$ and $\pi_i = j$. $N_{ij}=\sum_{t=1}^{r_i}\alpha_{ijt}$, and π_i is the parent attribute of X_i. We let ϕ_i denote a list of the unique parents of X_i as shown in D. If X_i has no parent, then we define ϕ_i as the list ϕ, where ϕ represents the empty set of parents. Then, $q_i = |\,\phi_i\,|$.

Using the metric shown by Equation (2), we compute the score of each BN. The heuristic greedy search algorithm is used to obtain the most optimal structure. Next, we assume an attribute as the root. Based on the obtained prior attribute ordering, we incrementally add parent attributes that can increase to the highest extent the probability of the resulting structure. Thus, we locally identify the most optimal structure of BN. We then determine the parent attributes of the attribute X_i. By repeating the procedure, we obtain the parent attributes of all attributes. Finally, the construction of BN is completed and named as the BN-IG system.

2.4. **Algorithm for constructing the BN-IG system.**

Input: dataset

Output: BN-IG system
Procedure:

1. Perform data preprocessing.
2. Compute the information gain. X_i and X_j are attribute variables for $i \neq j$ and $i, j = 1, \ldots, n$; and n is the number of attributes in the dataset. Each X_i and X_j has m cases; $k = 1, \ldots, m$, where m is the number of patients in the dataset. Then, our dataset can be represented as an $m \times n$ matrix, in which the (i, j)th component x_{ik} denotes the value of the i-th case and k-th attribute. The information gain is computed using

$$\begin{aligned} IG(X_i; X_j) &= H(X_i) - H(X_i \mid X_j) \\ &= -\sum_{k=1}^{m} P(X_i = x_{ik}) \log_2(P(X_i = x_{ik})) + \sum_{k=1}^{m} P(X_j = x_{jk}) H(X_i \mid X_j = x_{jk}). \end{aligned}$$

3. Use the "queue device" $P(IG(X_i; X_j) > \varepsilon)$ to rank these attributes.
4. Construct the BN.
   4.1. Repeat for each case and set initialized parameter as $\pi_i = \phi$:

$$P_{old} = f(X_i, \pi_i) = K_2(X_i, \pi_i); Flag = true.$$

   K_2 is the score determined using Equation (2). We use the K_2 score to guide the search for the optimal (with search-algorithm constraints) BN.
   4.2. Iterate for each $flag = true$ and $|\pi_i| < u$. Update $z = P_{red}(X_i) - \pi_i$ that maximizes $K_2(X_i, \pi_i \cup \{z\})$ and $P_{new} = f(X_i, \pi_i \cup \{z\})$. Here, $P_{red}(X_i)$ is the prior attribute ordering and u is the allowable maximum number of parent attributes (in our experiments, $u = 4$).
   4.3. If $(P_{new} > P_{old})$, then $P_{old} = P_{new}$ and $\pi_i = \pi_i \cup \{z\}$.

3. **Results.** The system is implemented using MATLAB software in an IBM computer with a 2.0 GHz processor and Windows XP operating system.

3.1. **Results of the CVD dataset.** The fivefold cross-validation approach is adopted to verify the robustness of the system. The dataset is randomly divided into five mutually exclusive and exhaustive groups. At each experiment, one group is selected as the test set and the other four groups are mixed together as the training set. The BN-IG system is constructed from the training set, and the performance is estimated using the corresponding test set.

The multivariate nonlinear associations among CVDs and related risk factors are shown in Figure 1. Given the states of attributes, the posterior probabilities of related CVDs can be computed from the BN. Then, the state of any risk factor can be predicted based on the probabilities. For example, Figure 1 shows that CVD (X_7) directly depends on age (X_1), gender (X_2), hypertension (X_3) and diabetes mellitus (X_4). In this case, we can calculate the probability of each subject $q = P(X_7 \mid X_1, X_2, X_3, X_4)$ based on the states of X_1, X_2, X_3 and X_4 from the BN in Figure 1. If $q \leq 0.5$, we can predict the state of the attribute CVD(X_7) as "absent"; otherwise, its state is "present". The results are consistent with those in literature [10,11].

By comparing the predicted value with the true value, the predictive accuracy can be obtained. For example, the predictive accuracy of jointly using X_1, X_2, X_3 and X_4 to predict X_7 is 0.76. If we separately use the four attributes, the accuracy of X_1, X_2, X_3 and X_4 is 0.58, 0.55, 0.72 and 0.70, respectively. Compared with the other groups of attributes such as X_1, X_2 and X_5, the combinations of X_1, X_2, X_3 and X_4 have the

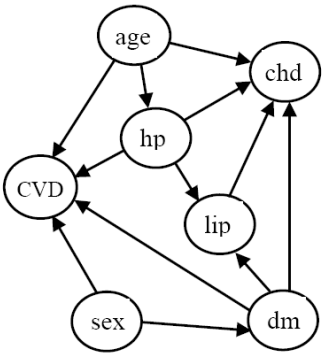

FIGURE 1. The BN of the CVD. hp, dm, chd, lip, and CVDs denote hypertension, diabetes mellitus, coronary heart disease, hyperlipemia, and cerebral infarction, respectively.

highest prediction. This result indicates that considering the states of the risk factors altogether is important in diagnosing CVDs.

Once the BN-IG system has been constructed, we can apply it for clinical CVD diagnosis. If we want to determine whether a patient is suffering from CVD (X_7), we can diagnose the possibility of X_1, X_2, X_3 and X_4. Thus, a diagnosis with relatively high accuracy in shorter time and less cost can be obtained.

The permitted maximum number of parent nodes is set as four in our experiment. Therefore, some relationships are weaker than the others and not displayed. Some research results have suggested that men have a much higher death rate from coronary heart disease than women (http://www.health.state.ny.us/nysdoh/heart/aboutchd.htm). We believe that the relationship between coronary heart disease and gender is weaker than the other associations in our research. Therefore, the relationship is not displayed in Figure 1.

3.2. **Comparison with other models.**

3.2.1. *Comparison with stepwise logistic regression based on predictive accuracy.* To validate further the effectiveness of the BN-IG system, we compare it with another multivariate analysis method: stepwise logistic regression. In this experiment, we apply stepwise logistic regression to the CVD dataset. This method has been previously used [6,12]. For the computation, we adopt multinomial regression to determine automatically which attribute to add or drop from the system in the software Statistical Package for the Social Sciences. Considering that logistic regression only supports a single dependent attribute, we view each risk factor as a dependent attribute and construct the corresponding regression system. In most cases, we achieve similar results with the proposed BN-IG system. For example, when we view X_5 as the dependent attribute, the stepwise logistic regression adds X_1, X_3, X_4 and X_6. No other attribute can be added to the system. Comparing these results with the results of the proposed system, stepwise logistic regression selects the same attribute relationships as the BN-IG system in most of experiments. The predictive accuracy of the two systems is shown in Table 2. We can observe that the predictive accuracy of the proposed BN-IG system is higher than that of the stepwise logistic regression system.

3.2.2. *Comparison with support vector machine (SVM) analysis based on the area under the receiver operating characteristic (ROC) curve (AUC).* To evaluate the performance of the proposed BN-IG system, we measure the performances of the BN-IG and SVM systems in terms of sensitivity, specificity, and classification accuracy. ROC analysis is conducted, and the AUC of the two systems is compared. AUC determination is a standard method

TABLE 2. The predictive accuracy between stepwise logistic regression and BN-IG

accuracy	hp	dm	chd	lip	CVD
Stepwise logistic regression	0.71	0.75	0.74	0.71	0.77
BN-IG	0.74	0.77	0.79	0.75	0.78

TABLE 3. The experimental results on CVD dataset

System	Sensitivity	Specificity	Accuracy	AUC
BN-IG	0.81	0.80	0.80	0.816
SVM	0.77	0.72	0.77	0.789

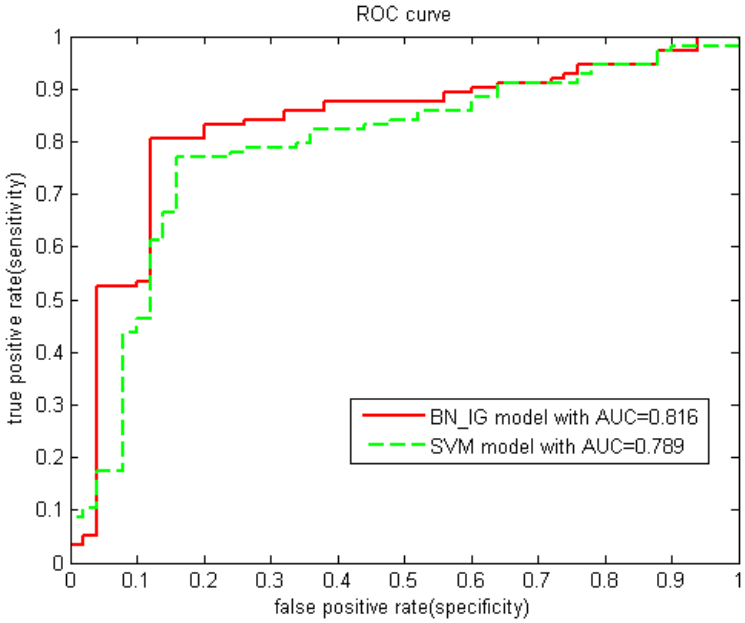

FIGURE 2. The ROC curve and AUC of BN-IG and SVM

of estimating the accuracy of a probabilistic pattern-recognition system [13]. Generally, larger AUC values indicate higher classifier performance.

For comparison with the SVM system, we implement the SVM system developed using the MATLAB and LIBSVM (URL: http://www.csie.ntu.edu.tw/_cjlin/libsvm) software packages. We linearly scale each attribute to the range $[0, +1]$ to avoid numerical problems. Radial basis function is selected as the kernel function. We apply grid search and cross-validation to identify the best SVM parameters. Grid search is performed within the range $\log_2 C \in \{-2, -1, \ldots, +9, +10\}$ and $\log_2 \gamma \in \{-10, -9, \ldots, +1, +2\}$. The parameter $\{C, \gamma\}$ that leads to the highest overall fivefold cross-validation classification accuracy in the training dataset is selected. Then, we use the best parameters to create an SVM classifier in the training dataset.

We randomly select X_6 as an example for comparing the performance of the two systems using the CVD dataset. We split the CVD dataset into training and testing sets at a 80% : 20% ratio. Table 3 lists the experimental results, and Figure 2 shows the ROC curve of the two systems. The BN-IG system is found to perform better than the SVM system using X_6 of the CVD experiment.

State-of-the art SVM implementations typically have a training time complexity that scales between $O(m)$ and $O(m^{2.3})$, where m denotes the number of training samples

TABLE 4. The description and experimental results on UCI datasets

Dataset	Number of Cases	Number of Attributes	Accuracy of SVM (%)	Accuracy of K2 (%)	Accuracy of BN-IG (%)
Iris	150	4	99.33	97.47	99.67
Monks	432	6	67.13	65.23	67.82
BUPA	345	6	58.55	58.03	59.71
Breast cancer	683	9	97.07	94.88	96.49
PID	768	8	77.73	74.48	76.17
Crx	653	15	66.77	64.32	64.47
Wine	138	13	71.74	67.39	72.46
Ionosphere	351	34	95.44	89.74	95.73

[14]. The complexity can be further scaled down to $O(m)$ with the use of a parallel mixture. However, these observations are only empirical and not based on theory. The overall complexity of BN-IG is $O(n^3)$. Although the complexity of BN-IG is higher than that of SVM, more functions are provided in BN-IG. Analysis reveals that the BN-IG system performs classification, supports probabilistic reasoning, and determines associations between attributes without multiple comparison problems.

3.2.3. *Comparison with SVM and K2 using the standard dataset.* To evaluate further the performance of the proposed BN-IG system, we test the system using the eight standard datasets from the UCI repository (http://www.ics.uci.edu/∼mlearn/MLRepository.html) and compare the results with SVM and the K2 algorithm. The specific procedures and parameters of SVM are similar to that in Section 3.2.2. The prior node ordering is randomly selected in the K2 algorithm. Other procedures and parameters of K2 are the same as that of BN-IG. These datasets are Iris, Monks, BUPA, Breast cancer, Pima Indians Diabetes, Crx, Wine and Ionosphere. The description and fivefold cross-validation experimental results using the UCI datasets are shown in Table 4. The accuracy of BN-IG is better than that of K2 algorithm. More nodes correspond to higher accuracies of two algorithms in most conditions. The accuracy using the five datasets is improved compared with that using SVM. The accuracy using SVM is slightly lower than that using the three datasets in the BN-IG system. The experimental results show that our system is superior to SVM in most conditions. The accuracy of the K2 algorithm is slightly lower than that of SVM, which results from the randomness of the node ordering. In addition, the BN-IG system and K2 algorithm have classification functions and probabilistic reasoning capabilities.

4. **Discussions.** By applying the BN-IG system to the CVD dataset, we can find nonlinear multivariate probabilistic associations among CVDs and related risk factors. For example, coronary heart disease directly depends on age, hypertension, diabetes mellitus, and hyperlipemia. We also find weaker associations of gender and CVDs with coronary heart disease, consistent with other reports [15]. Benner et al. [15] indicated that nearly all patients with coronary heart disease have prior exposure to at least one of the major coronary heart disease risk factors, including hypertension, hyperlipemia, cigarette use, and diabetes mellitus. Among them, hypertension and hyperlipemia are the two most common and readily modifiable coronary heart disease risk factors. Moreover, 85% of all coronary heart disease deaths occur in people aged 65 years or more. Coronary heart disease risk increases with age. Benner et al. [15] also reported that coronary heart disease is related to the color of skin, levels of income and education, as well as family history. However, these indicators are weakly associated with other risk factors of CVDs and thus not considered.

In the present study, we emphasize on the diagnosis of CVDs and associations among most risk factors of CVDs (http://www.health.state.ny.us/nysdoh/heart/aboutchd.htm).

We find that the attributes of age, hypertension, and diabetes mellitus are the strongest ones for diagnosing and predicting CVD (Figure 1). We observe that three nodes are directly associated with CVDs and related risk factors, consistent with previous reports [16-19]. For example, hypertension has an important function in CVDs and is the most important modifiable factor and second most important risk factor (after age) for hemorrhagic and ischemic stroke (CVD is the main disease of ischemic stroke) [16]. Diabetes mellitus is a risk factor of ischemic stroke [17]. Experimental data also suggest several gender differences in the risk of cerebral infarction (a CVD) in young patients [18]. Lina [19] used multivariate forward logistic regression analysis and found that the strongest predictors of CVD risk factors are age, gender, level of education, and length of residence, which are similar to our results. However, the level of education and length of residence are found to be the main risk factors, different from the present results. The discrepancy may be due to the differences in the method of analysis, data content, and race of subjects.

We construct the BN-IG system that can be used to determine how risk factors influence one another and quantify the extent of the influences. The BN-IG system also shows the result of interactions under clinical conditions. This method differs from univariate analysis, which focuses on specific regional effects. For example, in a comparison between CVD and normal groups, univariate analysis focuses on determining whether a specific factor such as hypertension has different levels between groups. By contrast, the BN-IG system examines the interactions among related risk factors such as age, hypertension, diabetes mellitus, and gender that commonly affect the diagnosis of CVD. We can also obtain the degree of influence of each factor on the diagnosis of CVD. Therefore, these results can be applied in clinical diagnosis, in which the risk of disease is based on age, gender, as well as extent of hypertension and diabetes mellitus of a clinical patient.

Furthermore, we can utilize the results as a guide for treating and preventing related CVDs. For example, we identify age, hypertension, and diabetes mellitus as the main risk factors of related CVDs. Thus, the focus of the treatment and prevention of related CVDs should be on hypertension and diabetes mellitus. Furthermore, we can also construct a BN-IG system for other illnesses to provide a rapid and general diagnosis.

Expert knowledge and experience can also be incorporated into the BN-IG system. For example, if an expert knows that a strong association exists between coronary heart disease and high blood pressure, we can develop an effective model that includes only the association between these two attributes. Expert knowledge on the probability distribution of an attribute can also be incorporated into the process of model generation. For example, if an expert believes that the probability of CVD being normal is 0.97, we can accordingly set the prior distribution of this feature.

This study has some limitations. The BN-IG system is generated using an optimization search algorithm in which the solution is an NP-hard problem. In the experiment, we adopt a greedy algorithm to search the most optimal network structures. The greedy search algorithm is an optimal algorithm. Therefore, the resultant network may not be a globally optimal result. In the future, a better optimal algorithm can be applied to obtain more stable results.

5. **Conclusions.** In this study, we develop a method using information gain technology to construct an optimized BN-IG system for a specific clinical CVD dataset. The BN-IG system can be directly used to diagnose and predict CVDs as well as reveal the complex, nonlinear multivariate associations among CVDs and related risk factors. We find the strongest factors of diagnosis and predictors for CVDs that can help diagnose and predict

this disease. This method can also be used to construct BN systems for other illnesses and provide functions for probabilistic inferences, which may contribute to effective diagnosis and prediction and significantly help patients and doctors.

Acknowledgements. This work was supported by the National Natural Science Foundation of China (No. 60971096) and the Liaoning Provincial Scientific Research Project (No. L2012381).

REFERENCES

[1] A. I. Hatzitolios, T. P. Didangelos, A. T. Zantidis, K. Tziomalos, G. A. Giannakoulas and D. T. Karamitsos, Diabetes mellitus and CeVD: Which are the actual data? *Journal of Diabetes and Its Complications*, vol.23, pp.283-296, 2009.

[2] X. Z. Zhu and Y. L. Gao, Risk factors and treatment of the CeVDs, *China Medical Herald*, vol.4, no.35, pp.154-155, 2007.

[3] P. B. Isabelle, P. Franck, E. Nathalie et al., Relationships between common polymorphisms of adenosine triphosphat e-binding cassette transporter A1 and high-density lipoprotein cholesterol and coronary heart disease in a population with type 2 diabetes mellitus, *Metabolism Clinical and Experimental*, vol.58, pp.74-79, 2009.

[4] T. Murase, M. Okubo, A. K. Michiyo et al., Impact of elevated serum lipoprotein (a) concentrations on the risk of coronary heart disease in patients with type 2 diabetes mellitus, *Metabolism Clinical and Experimental*, vol.57, no.6, pp.791-795, 2008.

[5] D. Y. Yeh, C. H. Cheng and Y. W. Chen, A predictive model for cerebrovascular disease using data mining, *Expert Systems with Applications*, vol.38, no.7, pp.8970-8977, 2011.

[6] R. Chen and E. H. Herskovits, Network analysis of mild cognitive impairment, *NeuroImage*, vol.29, no.4, pp.1252-1259, 2006.

[7] H. S. Park and S. B. Cho, Evolutionary attribute ordering in Bayesian networks for predicting the metabolic syndrome, *Expert Systems with Applications*, vol.39, pp.4240-4249, 2012.

[8] G. F. Cooper and E. H. Herskovits, A Bayesian method for the induction of probabilistic networks from data, *Machine Learning*, vol.9, no.4, pp.309-347, 1992.

[9] D. Heckerman, D. Geiger and D. M. Chickering, Learning Bayesian networks: The combination of knowledge and statistical data, *Machine Learning*, vol.20, no.3, pp.197-243, 1995.

[10] X. W. Yuan, Z. G. Du, D. Zhang et al., Whether chronic bronchitis is an independent risk factor for cerebral infarction in the elderly 1:1 case paired study, *Neural Regeneration Research*, vol.2, no.8, pp.502-505, 2007.

[11] X. Q. Zheng and H. Zhou, The pathophysiology of cerebral infarction in diabetes, *International Journal of CeVDs*, vol.15, no.3, 2007.

[12] A. T. Du, N. Schuff, D. Amend et al., Magnetic resonance imaging of the entorhinal cortex and hippocampus in mild cognitive impairment and Alzheimer's disease, *J. Neurol. Neurosurg Psychiatry*, vol.71, no.4, pp.441-447, 2001.

[13] J. H. Lin and P. J. Haug, Exploiting missing clinical data in Bayesian network modeling for predicting medical problems, *Journal of Biomedical Informatics*, vol.41, no.1, pp.1-14, 2008.

[14] J. C. Platt, Fast training of support vector machines using sequential minimal optimization, *Advances in Kernel Methods-Support Vector Learning*, Cambridge, MA, USA, pp.185-208, 1999.

[15] J. S. Benner, T. W. Smith, A. A. Petrilla et al., Estimated prevalence of uncontrolled hypertension and multiple cardiovascular risk factors and their associated risk of coronary heart disease in the United States, *Journal of the American Society of Hypertension*, vol.2, no.1, pp.44-53, 2008.

[16] F. Veglio, C. Paglieri, F. Rabbia et al., Hypertension and cerebrovascular damage, *Atherosclerosis*, vol.205, no.2, pp.331-341, 2009.

[17] Y. Sun and P. H. S. T. Matthias, Impact of diabetes mellitus (DM) on the health-care utilization and clinical outcomes of patients with stroke in Singapore, *Value in Health*, vol.12, no.S3, pp.S101-S105, 2009.

[18] B. Zhang, S. X. Pu, W. Z. Zhang et al., Sex differences in risk factors, etiology, and short-term outcome of cerebral infarction in young patients, *Atherosclerosis*, vol.216, no.2, pp.420-425, 2011.

[19] S. A. Lina, Cardiovascular disease risk factors among adult Australian-Lebanese in Melbourne, *International Journal of Research in Nursing*, vol.6, no.1, pp.1-7, 2010.