Jan. 2003, Vol.18, No.1, pp.97–101 J. Comput. Sci. & Technol.

Approach to the Correlation Discovery of Chinese Linguistic Parameters Based on Bayesian Method

WANG Wei (王 玮) and CAI LianHong (蔡莲红)

Department of Computer Science and Technology, Tsinghua University, Beijing 100084, P.R. China

E-mail: wangwei999@hotmail.com

Received May 22, 2001; revised March 26, 2002.

Abstract Bayesian approach is an important method in statistics. The Bayesian belief network is a powerful knowledge representation and reasoning tool under the conditions of uncertainty. It is a graphics model that encodes probabilistic relationships among variables of interest. In this paper, an approach to Bayesian network construction is given for discovering the Chinese linguistic parameter relationship in the corpus.

Keywords data mining, Bayesian approach, Chinese linguistic parameter

1 Introduction

With the development of the technology in speech processing, the Mandarin speech synthesis system has made great progress during the last few years, and has been used in various places successfully. But the result of it is still far away from the high naturalness compared with human for processing prosodic parameters. In recent years, some attempts have been made on the research of neural network for some western languages[1]. They resulted in noticeably better synthetic speech than the traditional rule-based approach. Not only the performance was improved but the system could also be easily configured for different persons by learning from existing databases automatically. But for Mandarin, it is still in primary status to process the prosody with the neural network. Mandarin is a tonal language. Much different to western languages, it is more complex. Linguistic parameter relationship influences the output of mandarin speech synthesis directly. Up to now, it is difficult to describe this relationship by a statistic method because we do not know prior knowledge. That is, we cannot conclude that among these parameters which are useful which are not. In the past years, this linguistic parameter relationship was represented by linguist experience. It is subjective but not objective. We need to find another approach to solving this problem, which is based on data-driven method entirely.

Data mining is for searching valuable information in large volumes of data. It is a cooperative effort of human and computers. Predictive data mining is for searching very strong patterns in big data that can generalize accurate future decisions. Data mining, also called knowledge discovery in database, is defined as the non-trivial process to identify valid, novel, potentially useful, and ultimately understanding patterns in data.

A Bayesian network is a graphical model for probabilistic relationship among a set of variables. Bayesian network has become a popular representation for encoding uncertain expert systems. There are numerous representations available for data mining, including rules bases, decision trees and artificial neural networks, classification, clustering, etc. There are at least four prospects[2]:

1) Bayesian networks can readily handle incomplete data sets. This correlation is not a problem for standard supervised learning techniques, all provided inputs are measured in every case. When one of the inputs is not observed, however, many models will produce an inaccurate prediction, because they do not encode the correlation between the input variables. Bayesian networks offer a natural way to encode such dependencies.

2) Bayesian networks allow one to learn about causal relationship, that is important for at least two reasons. The process is useful when we are trying to gain understanding about a problem domain. In addition, knowledge of causal relationships allows us to make predictions in the presence of interventions. The use of Bayesian networks helps to answer such questions even when no experiment about the effects of increased exposure is available.

3) Bayesian networks in conjunction with

Supported by the National Natural Science Foundation of China (Grant No.69875008).

Bayesian statistical techniques facilitate the combination of the domain knowledge and data. Anyone who has performed a real-world modeling task knows the importance of prior or domain knowledge, especially when data is scarce or expensive. The causal semantics of Bayesian networks makes the encoding of causal prior knowledge particularly straightforward. In addition, Bayesian networks encode the strength of causal relationships with probabilities. Consequently, prior knowledge and data can be combined with well-studied techniques of Bayesian statistics.

4) Bayesian methods in conjunction with Bayesian networks and other types of models offer an efficient and principled approach to avoiding the over fitting of data.

Summing up the above, we can draw the conclusion that Bayesian network is a common data-mining technique. In this paper, we give an approach to constructing Bayesian network and use Bayesian approach to discovering the relationship of Chinese linguistic parameters in the corpus. Discovering the relationship of linguistic parameters will help us to determine the importance of linguistic parameters in the group and simplify the input units of the training neural network effectively. The enormous neural network structure size may worsen convergence time and computational precision. It influences performance of the whole neural network directly.

This paper consists of five sections. In Section 2, we introduce Bayesian approach. In Section 3, we present an algorithm for Bayesian network construction. In Section 4, the Mandarin prosody influenced by the linguistic environment is analyzed. We use Bayesian approach to discover the relationship of linguistic parameters in the corpus. In Section 5, we conclude this paper and point out the future work.

2 Bayesian Approach

In the Bayesian approach, we use the causal Markov assumption to look for structures that fit conditional-independence constraints. Furthermore, because the Bayesian approach uses a probabilistic framework, we no longer need to make decisions on individual-independence facts. We can compute the probability that the independencies associated with an entire casual structure are true.

Suppose our problem domain consists of variables $X = \{X_1, \ldots, X_N\}$. We have some data $D = \{x_1, \ldots, x_N\}$ which is a random sample from some unknown probability distribution for variables domain X. Here, we assume that each case x in D consists of an observation of all the variables in X, and the structure of this causal model is a directed-acyclic graph that encodes conditional independencies via the causal Markov assumption.

Definition 1. *We define a discrete variable M whose states m correspond to the possible true models, and encode our uncertainty about M with the probability distribution $p(m)$.*

Definition 2. *For each model structure m, we define a continuous vector-valued variable θ_m, whose values θ_m correspond to the possible true parameters.*

Definition 3. *We encode uncertainty about θ_m using the probability density function $p(\theta_m|m)$.*

Given a random sample D, we compute the posterior distribution for each m and θ_m using Bayesian rules:

$$p(m|D) = \frac{p(m)p(D|m)}{\sum_{m'} p(m')p(D|m')} \tag{1}$$

$$p(\theta_m|D,m) = \frac{p(\theta_m|m)p(D|\theta_m,m)}{p(D|m)} \tag{2}$$

where $p(D|m) = \int p(D|\theta_m,m)p(\theta_m|m)d\theta_m$ is called the marginal likelihood. Given some hypothesis of interest h, we determined the probability that h is true given data D by averaging over all possible models and their parameters:

$$p(h|D) = \sum_m p(m|D)p(h|D,m) \tag{3}$$

$$p(h|D,m) = \int p(h|\theta_m,m)p(\theta_m|D,m)d\theta_m \tag{4}$$

is the likelihood of a model. Under certain assumptions, these computations can be done efficiently and in a closed form. One assumption is that the likelihood term $p(x|\theta_m,m)$ factorizes as follows:

$$p(x|\theta_m,m) = \prod_{i=1}^{n} p(x_i|pa_i,\theta_i,m) \tag{5}$$

where each local likelihood $p(x_i|pa_i,\theta_i,m)$ is in the exponential family, pa_i denotes the configuration of the variables corresponding to the parents of node x_i, θ_i denotes the set of parameters associated with the local likelihood for variable x_i. Once each variable $x_i \in X$ is discrete, having r_i possible values $x_i^1, \ldots, x_i^{r_i}$, and each local likelihoods is a collection of multinomial distributions, one distribution for each configuration of pa_i, that is:

$$p(x_i^k|pa_i^j,\theta_i,m) = \theta_{ijk} > 0 \tag{6}$$

where $pa_i^1, \ldots, pa_i^{q_i}$ $\left(q_i = \prod_{x_i \in pa_i} r_i\right)$ denote the configurations of pa_i and $\theta_i = ((\theta_{ijk})_{k=2}^{r_i})_{j=1}^{q_i}$ are parameters. The parameter θ_{ij1} is given as follows:

$$\theta_{ij1} = 1 - \sum_{k=2}^{r_i} \theta_{ijk} \tag{7}$$

Definition 4. *The vector of parameters is defined as* $\boldsymbol{\theta}_{ij} = (\theta_{ij2}, \ldots, \theta_{ijr_i})$, $\forall i, j$.

Given a random sample D that contains no missing observations, the parameters remain independent:

$$p(\theta_m | D, m) = \prod_{i=1}^{n} \prod_{j=1}^{q_i} p(\boldsymbol{\theta}_{ij} | D, m) \tag{8}$$

We can update each vector of parameter $\boldsymbol{\theta}_{ij}$ independently. Assuming each vector $\boldsymbol{\theta}_{ij}$ has Dirichlet distribution $Dir(\boldsymbol{\theta}_{ij} | \alpha_{ij1}, \ldots, \alpha_{ijr_i})$, we obtain the posterior distribution for the parameters:

$$p(\boldsymbol{\theta}_{ij} | D, m) = Dir(\boldsymbol{\theta}_{ij} | \alpha_{ij1} + N_{ij1}, \ldots, \alpha_{ijr_i} + N_{ijr_i}) \tag{9}$$

where N_{ijk} is the number of cases in D in which $X_i = x_i^k$ and $Pa_i = pa_i^k$. The collection of counts N_{ijk} is sufficient statistics of the data for the model m. We obtain the marginal likelihood:

$$p(D|m) = \prod_{i=1}^{n} \prod_{j=1}^{q_i} \frac{\Gamma(\alpha_{ij})}{\Gamma(\alpha_{ij} + N_{ij})} \prod_{k=1}^{r_i} \frac{\Gamma(\alpha_{ijk} + N_{ijk})}{\Gamma(\alpha_{ijk})} \tag{10}$$

where $\alpha_{ij} = \sum_{k=1}^{r_i} \alpha_{ijk}$ and $N_{ij} = \sum_{k=1}^{r_i} N_{ijk}$. We use (1) and (10) to compute the posterior probability $p(m|D)$.

3 Algorithm for Bayesian Network Construction

Initiate a graph $G(V, E)$ where $V = \{$all the nodes of a data set$\}$, $E = \{\}$. Initiate two empty ordered set $S = \emptyset$, $R = \emptyset$[3].

1. For each pair of nodes (x_i, x_j) where $x_i, x_j \in V$, compute mutual information $I(x_i, x_j)$ of two nodes x_i, x_j using the following equation:

$$I(x_i, x_j) = \sum_{x_i, x_j} P(x_i, x_j) \log \frac{P(x_i, x_j)}{P(x_i)P(x_j)} \tag{11}$$

Conditional mutual information is defined as:

$$I(x_i, x_j | c) = \sum_{x_i, x_j} P(x_i, x_j, c) \log \frac{P(x_i, x_j | c)}{P(x_i|c)P(x_j|c)} \tag{12}$$

where c is a set of nodes. If $I(x_i, x_j) > \varepsilon$ (ε is a certain small value) then sort them by their mutual information from large to small and put them into an ordered set S.

2. Get the first two pairs of nodes in S and remove them from S. Then add the corresponding arcs to E.

3. Get the first pair of nodes remained in S and remove it from S. If there is no open path between the two nodes, add the corresponding arc to E. Otherwise, add the pair of nodes to the end of an ordered set R.

4. Repeat Step 3 until $S = \emptyset$.

5. Get the first pair of nodes in R and remove it from R.

6. Find a block set with each open path between these two nodes by a set of minimum number of nodes. If these two nodes are still dependent on each other given the block set, connect them by an arc.

7. Go to Step 5 until $R = \emptyset$.

8. For each arc in E, if there are open paths between the two nodes besides this arc, remove this arc from E temporarily and find a block set with each open path between these two nodes by a set of minimum number of nodes. Conduct a conditional independence test under the condition of the block set. If $P(x_i, x_j) = P(x_i)P(x_j)$, add this arc back to E, otherwise remove this arc.

Algorithm complexity Analysis:

Suppose a data set has N attributes, the maximum number of possible values of any attribute is r, and an attribute may have k parents at most. We give the complexity as follows.

Computing mutual information between any two nodes needs N^2 mutual information computations. By (11), each computation of mutual information requires $O(r^2)$ times of basic operations such as logarithm, multiplication and division. Sorting the node pairs can be finished in $O(N \log N)$ steps by using quick sort algorithm. The time complexity of this phase on basic operation is $O(N^2 r^2)$. By (12), each conditional independent test requires at most $O(r^{k+2})$ basic operations. The complexity of this phase on basic operations is $O(N^2 r^{k+2})$. In the worst case, it requires basic operations $O(N^2 r^N)$ times. Removing arc phase has the same complexity as adding arc.

The algorithm requires conditional independence tests of $O(N^2)$ complexity. The time complexity on basic operations is $O(N^2 r^{k+2})$. In the worst case, when all the conditional independence tests require condition-sets on all the other nodes, the time complexity on basic operation is $O(N^2 r^N)$.

4 Correlation Discovery of Chinese Linguistic Parameters

In this paper, we establish several models to display these relationships according to different databases. We apply Bayesian approach to complete and incomplete databases in different assumptions. The linguistic parameter databases consist of five groups: the current syllabic parameters $\boldsymbol{C}$, the preceding syllabic parameters $\boldsymbol{L}$, the next syllabic parameters $\boldsymbol{N}$, the phrasal parameters $\boldsymbol{W}$, and the sentence parameters $\boldsymbol{S}$.

Fig.1 to Fig.4 show a Bayesian network built from a database consisting of 1840 records. Fig.5 to Fig.8 show a Bayesian network built from a database consisting of 1000 records. Fig.9 to Fig.12 show a Bayesian network built from a database consisting of 500 records[4,5].

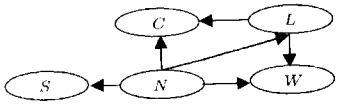

Fig.1. Complete database.

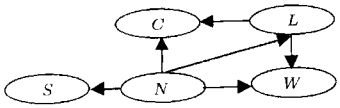

Fig.2. 1% incomplete data database.

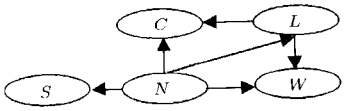

Fig.3. 2% incomplete data database.

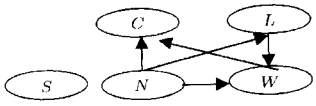

Fig.4. 5% incomplete data database.

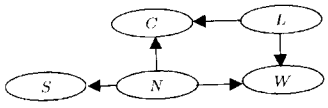

Fig.5. Complete database.

The results show that Bayesian approach can be used to establish models for databases with incomplete data. The smaller the data scale, the more accurate the model. When databases miss less data, the model depending on incomplete database is the same as those depending on complete database. If the missing data is more than a given threshold value, for example 5%, errors will occur. The results demonstrate that Bayesian approach is an effective tool for constructing network from databases.

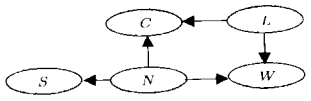

Fig.6. 1% incomplete data database.

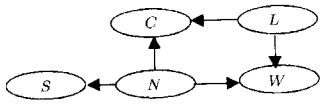

Fig.7. 2% incomplete data database.

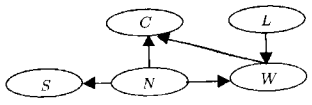

Fig.8. 5% incomplete data database.

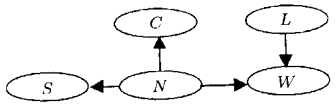

Fig.9. Complete database.

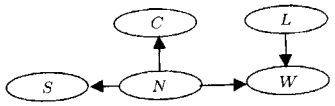

Fig.10. 1% incomplete data database.

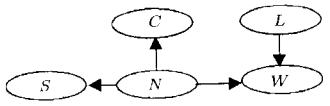

Fig.11. 2% incomplete data database.

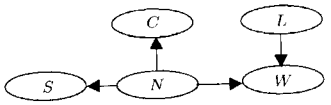

Fig.12. 5% incomplete data database.

Then we select 17 linguistic parameters in a speech synthesis system[6]. These parameters includes the initial type c_1, the final type c_2, the tone type c_3, the position in the current word c_4, the tight degree to the preceding syllable c_5, the tight degree to the next syllable c_6, the final type l_1, the tone type l_2, the initial type n_1, the tone type n_2, a the number of syllables in a group w_1, the position in a sentence w_2, the stress degree w_3, the distance

to the previous stress w_4, the distance to the next stress w_5, the sentence type s_1 and the number of phrases in a sentence s_2.

One method is calculating arithmetic average value of these parameters in different speech environment. These parameters can be divided into four levels by linguist experiences. The result is shown in Table 1.

Table 1. Arithmetic Average Value in Different Speech Environments

Linguistic parameters	Level 1	Level 2	Level 3	Level 4
c_1	4	3	4	3
c_2	4	3	4	1
c_3	5	2	3	3
c_4	3	3	4	4
c_5	2	0	0	0
c_6	1	1	1	2
l_1	0	0	0	1
l_2	0	2	3	4
n_1	8	5	6	6
n_2	9	8	8	8
w_1	8	8	4	5
w_2	9	8	7	9
w_3	3	4	3	3
w_4	1	2	2	1
w_5	4	3	3	3
s_1	1	2	1	1
s_2	3	4	5	3

In Table 1 the number values represent different voice features at different levels. It is difficult to find a relationship from the above table because all parameters change from one speech environment to another speech environment. That is, the arithmetic average value cannot reflect the relationship effectively. By Bayesian method, we can obtain the following graph as shown in Fig.13.

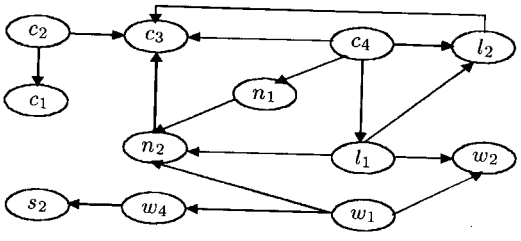

Fig.13. Graph of linguistic parameters in the corpus.

From Fig.13, we conclude that parameters c_5, c_6, w_3, w_5, s_1 have no relationship to other parameters. We need to remove them from the linguistic parameter group. The remained parameters will be input to the next neural network for training. The Bayesian graph model reflects the importance of linguistic parameters in the group.

5 Conclusion

This paper describes Bayesian network and Bayesian approach. We present an algorithm to construct Bayesian network structure for discovering Chinese linguistic parameter relationship in corpus. As a result, we obtain the relationship of 17 linguistic parameters. It helps us to simplify the next neural network structure and display the data relationship by graph models. But the missing data number has certain influence on the Bayesian network construction result. How to reduce the sensitivity of Bayesian network is our future research work.

References

[1] Karaali O *et al.* Text-to-speech conversion with neural networks: A recurrent TDNN approach. In *Proc. 5th European Conference on Speech Communication and Technology,* Rhodes, Greece, September 22–25, 1997, pp.561–564.

[2] David Heckerman. Bayesian networks for data mining. *Data Mining and Knowledge Discovery,* 1997, 1: 79–119.

[3] Cheng J, Bell D A, Liu W. An algorithm for Bayesian belief network construction form data. In *Proc. 6th International Workshop on Artificial Intelligence and Statistics,* Florida, USA, January 4–7, 1997.

[4] Wang Wei, Chen Enhong, Wang Xufa. Based on Bayesian approach for data mining. *Mini-Micro Systems,* 2000, 21(7): 703–705.

[5] Wang wei, Cai Lianhong. Study of determining Bayesian network topology structures. *Mini-Micro System,* 2002, 23(4): 435–437.

[6] Tao Jianhua, Cai Lianhong. A neural-network-based prosodic model of Mandarin TTS system. *Journal of Acoustics,* 2001, 26(1): 67–72.

[7] David Heckerman, Christopher Meek, Gregory Cooper. A Bayesian Approach to Causal Discovery, Technical Report MSR-TR-97-05.

[8] Gregory F Cooper, Edward Herskovits. A Bayesian method for the induction of probabilistic networks from data. *Marchine Learning,* 1992, 9: 309–347.

WANG Wei was born in 1973. He received the B.S. and M.S. degrees in electronic engineering and information science from Anhui University in 1995 and 1998 respectively, and the Ph.D. degree from the Department of Computer Science and Technology, University of Science and Technology of China in 2000. He is now a postdoctoral researcher at Tsinghua University. His research areas include multimedia information processing, data mining and machine learning.

CAI LianHong was born is 1945. She graduated from the Department of Automatic Control Engineering, Tsinghua University in 1970. She is now a professor at the Department of Computer Science and Technology, Tsinghua University. Her current research interests include speech synthesis, multimedia information processing.