Journal of Physics: Conference Series

An Improved K2 Algorithm for Symptom-Disease Bayesian Network

An Improved K2 Algorithm for Symptom-Disease Bayesian Network

Ying Yu, Wenwen Yang, Hao Huang, Zheng Yang and Jianye Lu

No.333 Nanchen Road, Baoshan District, Shanghai, China
Department of Mechanical and Electrical Engineering and Automation, Shanghai University
Email: yangwenwen222@163.com

Abstract. This paper establishes a Bayesian network for pre-diagnosis based on symptom-disease knowledge from clinical data. In order to solve the local optimal problem, the idea of simulated annealing algorithm is introduced into an improved K2 algorithm, and the BDE scoring function is used to determine the optimal network structure of the Bayesian network. performance evaluation indexes are used to evaluate the performance of the network. The experimental results show that the improved K2 algorithm outperformed the K2 algorithm.

1. Introduction

The clinical data contains a lot of useful knowledge, so relative clinical knowledge extraction has become a hot topic in academicals in recent years. This paper builds a Bayesian network based on acquired symptom-disease knowledge from clinical text data.

K2 algorithm was firstly proposed by Herskovits and Cooper in 1992[1]. Liu et al. proposed an improved K2 algorithm combining mutual information and ant colony algorithm[2], so that the problem that K2 algorithm relies on prior knowledge could be solved, and the search space of ant colony algorithm is reduced. Wei et al. proposed an improved K2 algorithm based on conditional mutual information and probability jump mechanism[3]. Li proposed an improved K2 algorithm based on particle swarm optimization algorithm[4].

Inspired by the idea of simulated annealing algorithm, we proposed a simulated annealing alike K2 algorithm to construct Bayesian network, and the network is evaluated by a BDE scoring function.

2. A Simulated Annealing Alike K2 Algorithm to Construct Bayesian Network

2.1. BDE Scoring Function

Scoring function is used to evaluate the fitting degree between network structure and data set, which is an important index to judge the structure of Bayesian network [5]. At present, the commonly used scoring functions are Bayesian Dirichlet-Likelihood Equivalence (BDE)[6], Minimum Description Length (MDL)[7], Akaike Information Criterion (AIC)[8] and Bayesian Information Criterion (BIC)[9]. BDE scoring function is one of the earliest scoring functions for judging the Bayesian network data fit [10].

The formula of the BDE score is shown in (1):

$$p_{BDE}(D,B_S^h)=p(B_S^h)\prod_{i=1}^{n}\prod_{j=1}^{qi}\frac{\Gamma(a_{ij})}{\Gamma(a_{ij}+N_{ij})}\prod_{k=1}^{ri}\frac{\Gamma(a_{ijk}+N_{ijk})}{\Gamma(a_{ijk})} \tag{1}$$

Where, n is the number of variables, q_i is the number of configurations of the i-th variable X_i and the set of parent nodes $P_a(X_i)$. N_{ijk} is the sample number of variable X_i in sample data D when its parent node set $P_a(X_i)$ takes the j -th configuration and the variable state is k . N_{ij} is the sum of all state values of N_{ijk} ,that is $N_{ij} = \sum_{k=1}^{r_i} N_{ijk}$. α_{ijk} is a hyper-parameter of Dirichlet distribution in Bayesian network parameters, $\alpha_{ij} = \sum_{k=1}^{ri} \alpha_{ijk}$, $\Gamma(n)$ is gamma function, for any $n(n \in \mathbf{N})$, $\Gamma(n) = (n-1)!$.

2.2. The Simulated Annealing Alike K2 Algorithm

K2 algorithm needs to know the node order in advance, otherwise it is easy to fall into local optimization during network construction[11]. In order to solve this problem, this paper proposes an improved K2 algorithm by introducing the idea of simulated annealing into the process of network construction.

By integrating the idea of simulated annealing algorithm with some changes, this paper proposes an improved K2 algorithm. The specific changes are as follows:

(1) Stack for reserving Solutions

Simulated annealing algorithm retains only one feasible solution in one iteration, but the improved K2 algorithm builds a stack for reserving multiple candidate optimal solutions. The stack not only holds the optimal solution, but also some suboptimal solution with a certain probability. It is possible to reserve some temporal suboptimal solution which may convert into an optimal solution in the following iterations.

(2) Weak-to-strong mode

In the new iterative process, with the addition of a new node, if the score of some solution in stack is better than the current optimal score, the weak-to-strong mode would start: the candidate solution would be popped from the stack and replace the original optimal solution, and the original optimal one would be pushed into the stack.

The process of the improved K2 algorithm is as follows:

```
Input:
$\lambda$ : a set of feature words $\{\tau_0, \tau_1, \cdots, \tau_n\}$
$\mu$ : maximum number of parent nodes
Output: Bayesian Network $B_s$ with the largest score $V$

$B_s$ is a borderless graph consisting of symptom word node $\tau_0, \tau_1, \cdots, \tau_n$ and target disease $Y$.
$\pi_i \leftarrow \phi$
$\tau_{opt} \leftarrow \tau_0, V_{opt} \leftarrow score(Y, \pi_j \cup \{\tau_0\})$ ( $score(Y, \pi_j \bigcup \{\tau_j\})$ represents a network score consisting of the target disease $Y$ and a set of feature words $\pi_j$ containing $\tau_j$ .)
for $i = 1$ to $\mu$
   flag $\leftarrow$ false
    for $j = 0$ to $n$
      $V_j \leftarrow \underset{\tau_j \notin \pi_i}{score}(Y, \pi_i \cup \{\tau_j\})$
        if $(p(Y \mid \tau_j) > \theta_1)$ and $(p(\tau_j \mid \overline{Y}) < \theta_2)$ ( $\theta_1, \theta_2$ are threshold)
      if ($V_j < V_{opt}$)
        if ($\left|V_{opt} - V_j\right| < threshold$)($threshold$ is the given threshold)
          if ( $random(0,1) < \exp(\frac{V_j\text{-}V_{opt}}{K})$ ) ( $K$ represents a given coefficient, $K > 0$, increase as the difference between $V_{opt}$ and $V_j$ increases)
            $\text{stack}_i.\text{push}(\pi_i \cup \{\tau_j\})$ ( $\text{stack}_i$ is a stack that stores the temporary solution of the layer $i$ )
          end if
        end if
      else
        $V_{opt} \leftarrow V_j, \pi_{opt} \leftarrow \pi_i \cup \{\tau_j\}$
        flag $\leftarrow$ true
          while ( $\text{stack}_{i-1} \neq \phi$ )
            $\pi_{tmp} \leftarrow \text{stack}_{i-1}.pop()$
            $V_{tmp} \leftarrow \underset{\tau_j \notin \pi_{tmp}}{score}(Y, \pi_{tmp} \cup \{\tau_j\})$
              if ($V_j < V_{tmp}$)
                if ($\left|V_{tmp} - V_j\right| < threshold$) and ($random(0,1) < \exp(\frac{V_j\text{-}V_{tmp}}{K})$))
                  $\text{stack}_i.\text{push}(\pi_i \cup \{\tau_j\})$
                  $V_{opt} \leftarrow V_{tmp}, \pi_{opt} \leftarrow \pi_{tmp} \cup \{\tau_j\}$
                end if
              else
                if ($\left|V_{tmp} - V_j\right| < threshold$) and ($random(0,1) < \exp(\frac{V_{tmp} - V_j}{K})$))
                  $\text{stack}_i.\text{push}(\pi_{tmp} \cup \{\tau_j\})$
                end if
              end if
          end while
      end if
     end if
    end for
    if (flag=false)and($\pi_j \neq \phi$)
      break
   else
      Add an edge $\tau_{opt} \rightarrow Y$ to $B_s$
   end if
 end for
output $B_s$
```

3. Numerical Example

About 4,000 records of gastrointestinal diseases were extracted, where 2,000 records are training data and 2,000 records are test data. Take the disease "gastritis" as an example to construct its network. The specific process is as follows:

(1) Symptom words

About 12 symptoms words are to be chosen as symptoms of "gastritis", with the notations shown in Table 1:

Table 1. Symbol table of some feature words

Feature words	Symbol
stomach ache/stomachache/pain	C_1
gastric acid/acid reflux	C_2
stomach distension/stomach swelling	C_3
stomach fever/stomach heat/heartburn	C_4
vomiting/retching	C_5
loss of appetite/unwillingness/inability to eat	C_6
diarrhea/ loose stool	C_7
belly pain/stomachache/abdominal pain/abdominal pain	C_8
stool does not form/loose stool	C_9
left lower abdomen/left abdomen	C_{10}
hematochezia/purulent blood/hemorrhage after defecation/bloody stool	C_{11}
fever/high fever	C_{12}

(2) Bayesian Network Construction with the Improved K2 Algorithm

The optimal network BDE score obtained by the K2 algorithm and the improved K2 algorithm for "gastritis" disease are shown in Figure 1:

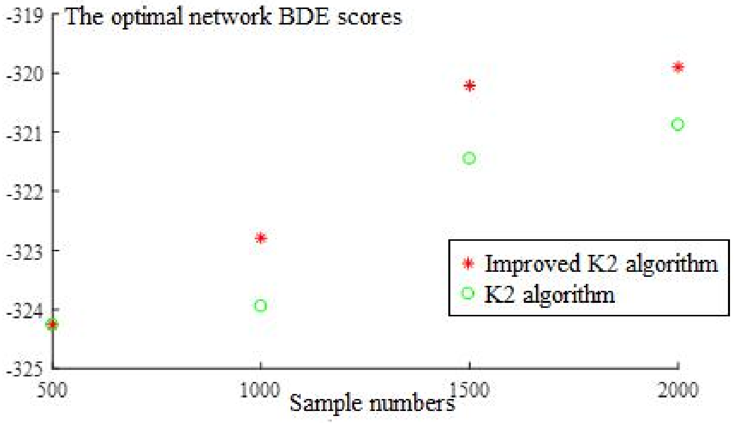

Figure 1. BDE scores of K2 and improved K2 algorithm for "gastritis" disease

It can be seen from Fig.1, with the increase of sample size, BDE score function shows an upward trend, and the improved K2 algorithm can obtain a network structure with better BDE scores in most cases,

The parent nodes and specific scores of the network structure under different sample sizes are shown in Table 2:

Table 2. Optimal Networks for " Gastritis" with K2 and improved K2 Algorithm

Sample size	K2 algorithm		Improved K2 algorithm	
	BDE sore	Parent node	BDE sore	Parent node
500	-323.6960	C_1&sour&hiccup	-323.6960	C_1&hiccup&sour
1000	-323.6338	C_1&sour&hiccup	-322.7797	C_1&C_2&C_4
1500	-321.4953	C_1&sour&C_4	-320.0310	C_1&C_2&C_4
2000	-320.8789	C_1&sour&C_4	-319.6585	C_1&C_2&C_4

As shown in Fig.1 and Table 2, the improved K2 algorithm outperforms the K2 algorithm.

The performances of K2 algorithm and improve K2 algorithm for "acute enteritis" is shown in Fig. 2:

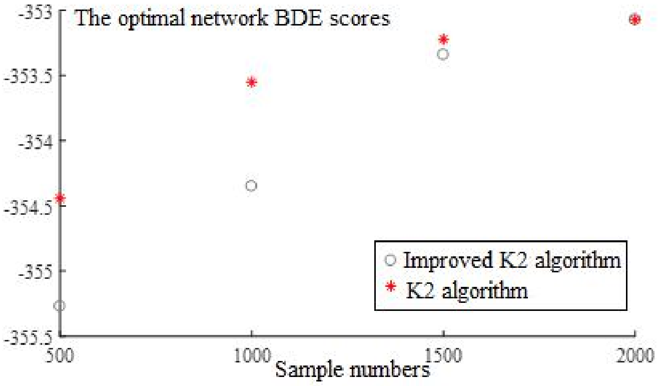

Figure 2. BDE scores of K2 and improved K2 algorithm for "acute enteritis" disease

From Fig. 1 and Fig. 2, it can be seen that the optimal BDE scores of the improved K2 algorithm are better than the optimal BDE scores of the original K2 algorithm.

(3) Comparison with evaluation indexes

Applying the constructed Bayesian network for diagnosis on test sample, and the comparative results are shown in Table 3.

Table 3. Results for two diseases

Disease	Gastritis		Acute enteritis	
Algorithm	K2	Improved K2	K2	Improved K2
Parent node	C_1 &acid	C_1&C_2&C_4	C_7&C_5&C_8	C_7&C_5&C_8

	vomitting & C_4			
TP	245	241	117	117
FN	189	193	291	291
FP	210	197	108	108
TN	1356	1369	1419	1419
Precision	80.05%	80.5%	79.38%	79.38%
Accuracy	53.85%	55.02%	52%	52%
Recall	56.45%	55.53%	28.68%	28.68%

In summary, among the two diseases, the performance index of using the improved K2 algorithm to construct the optimal network is better than or equal to the optimal network constructed by the traditional K2 algorithm, so the effectiveness of the improved algorithm is confirmed.

4. Conclusion

This paper proposes an improved K2 algorithm by incorporating the idea of simulated annealing algorithm. The experimental results show that the simulated annealing alike K2 algorithm can build a more robust Bayesian network. By extracting knowledge from clinical text data, the constructed Bayesian model can be used to assist pre-diagnosis in clinic and has practical significance.

5. References

[1] Cooper G F, Herskovits E. A Bayesian method for the induction of probabilistic networks from data[J]. Machine Learning, 1992, 9(4):309-347.

[2] Liu Hao, Sun Meiting, Li Lei, et al. Study on Bayesian network structure learning algorithm based on ant colony node order optimization [J]. Chinese Journal of Scientific Instrument, 2017(1).

[3] Wei Zhongqiang, Xu Hongwei, Li Wen, et al. Bayesian network structure learning algorithm based on conditional mutual information and probabilistic jumping mechanism[J]. Computer Science.

[4] Li Dongling. A Bayesian networks structure learning method based on particle swarm optimization modeling[J]. Journal of Computer Applications and Software, 2014(11):178-182.

[5] Dong Liyan.Research of Application Foundation on Bayesian Networks [D].Jilin University.

[6] Heckerman D, Geiger D, Chickering D M . Learning Bayesian networks: The combination of knowledge and statistical data[J]. Machine Learning, 1995, 20(3):197-243.

[7] Lam W, Bacchus F. Learning Bayesian networks: An approach based on the MDL principle [J]. Computational Intelligence, 1994, 10(3):269-293.

[8] Akaike H. A Bayesian Analysis of the Minimum AIC Procedure[J]. Annals of the Institute of Statistical Mathematics, 1978, 30(1):9-14.

[9] Burnham K P, Anderson D R. Multimodel Inference: understanding AIC and BIC in Model Selection[J]. Sociological Methods & Research, 2004, 33(33):261-304.

[10] An Ning, Teng Yue, Yang Jiaoyun, et al. Bayesian network structure learning method based on causal effect[J]. Journal of Computer Applications, 2018, 35(12): 95-99.

[11] Fu Ziyang. Study in the threat of violent terrorist activities with Bayesian network [D]. 2016.