Ontology-driven Bayesian Network Model for Semantic Expression

Alaa Aldeen Ismail Abuazab1,*, Harihodin Bin Selamat1, Rasimah Bt Che Mohd Yusoff1 and Ahmed N. Abdalla2

1Faculty of Advanced Informatics School, Universiti Teknologi, Johor, Malaysia
2Faculty Electronic Engineering, Huaiyin Institute of Technology, Jiangsu, China

*Corresponding author

Keywords: Bayesian network model, Medical diagnosis, Blood pressure.

Abstract. In this paper, ontology-driven Bayesian network model is proposed using the semantic ontology knowledge base, which automatically transforms the entities of the ontology into the Bayesian network model. Brief detail of the advantage of Bayesian network is applied for solving the uncertain and non-complete information. The hypertensive ontology is constructed to prove the validity of the model. The medical diagnosis algorithm based on the ontology-driven Bayesian network model to assist the NETICA Application Programming Interface (API). The proposed model used to realize the mapping between ontology and Bayesian network, and the different probability of the condition is entered to obtain the probability that the patient is suffering from high blood pressure. The experimental results show that the model is correct and feasible, and it has good universality and portability in medical diagnosis.

Introduction

The electronic health record (EHR) is a powerful repository of patient information that can be leveraged to build applications that benefit the clinical community such as Disease Prediction. Understanding and extracting information from EHRs enables reasoning with clinical variables and supports decision making [1-3]. Electronic Health records are store the patient information as data coded in structured format, as well as in the form of free text for clinical documentation. Structured data typically contains patient demographics information, patient birth and death information, lab test information, encounters, and at times procedures and diagnosis lists. Unstructured data includes free text clinical documentation that correspond to different encounters generated at various points of visit, including admission notes, history and physical reports, discharge summaries, radiology reports, and pathology reports. For overall information, the extracting information from EHR could be used to prediction disease module and computing the patients’ alert risk index and alert prediction index [4].

There are many of uncertainty in medical diagnostic operations, which are derived from incomplete information or unreliability of knowledge [5-7]. The ontology uses to solve the knowledge management and the Bayesian network approach to deal with the uncertainty of the diagnosis. Ontology as a knowledge base, can be domain knowledge expressed as a machine readable form; can be based on semantics to express a complex field of organizational structure, but it cannot handle the knowledge of uncertainty. The Bayesian network is widely used in obtaining the reasoning of uncertain knowledge by obtaining the confidence of uncertain knowledge [7]. The ontology contains the entity class and the object attribute and the relation attribute. Through a large number of instances, the ontology can depend on the object attribute to form the instance structure. The Bayesian network has its own variables, and these variables can form a relationship through dependencies. Therefore, it can transform the ontology and Bayesian networks, map the ontology instances into Bayesian network variables, and map the object relations between ontology entities into the dependencies between Bayesian network variables. The Bayesian network is a graphical model of probabilistic knowledge representation that is widely used for uncertainty [8]. The knowledge gained by people is subjective, random, ambiguous, and ambiguity. The Bayesian network expresses the uncertainty of things in the
real world in a reasonable way through probability. At the same time, to show uncertain knowledge, making the Bayesian network in different environments to be applied, especially in the direction of reasoning. Bayesian basic structure for the directional acyclic graph, which is responsible for the application of variables and variables between the relationships between the encoding, the associated digital part and the joint probability distribution associated coding. Bayesian network model which can structure and digital information modeling.

The innovation of this paper is to combine the advantages of ontology and Bayesian network to create an ontology-driven Bayesian network model to express the knowledge of ontology using Bayesian network to express its probability, by providing the patient with the probability of illness, the diagnosis of medical diseases to provide a reference. Then we construct the Bayesian network model from the body of hypertension on the basis of NETICA API, and finally use the Bayesian network to infer the patients with hypertension Probability.

Knowledge of Medical Field

At present, many scholars have studied the ontology and given its definition [9-10]. In essence, the ontology provides a well-defined knowledge. This knowledge can improve the accuracy and efficiency of communicating information with each other, resulting in other benefits such as operation, reuse, and sharing. The specific development process is as follows.

The Goal and Scope of the Ontology

In order to clearly describe the concept of medical field and the relationship between concepts, the medical field ontology constructed should not only meet the needs of the medical field, but also should have strong semantic reasoning, so the construction of the ontology requires the participation of occupational physicians.

Build the Ontology

First, the key concepts in the field should be described before constructing the ontology of the domain, and Fig.1 depicts the key concepts defined in the medical ontology. In the ontology, the class is the expression of the concept, and in this particular reference to the two main Figure 1 medical ontology class relationship, that is, the cause (Cause) and the results (Effect).

The subclasses of the cause are the environment and the Heredity. The subclasses of the results are Disease, Symptom, Sign, Test. The second step is to determine the relationship between the attributes of the concept. The two main attributes of this paper, as shown in Fig.2.

Any cause of at least one result, the implementation is Cause has Effect (inverse relationship is Effect of Cause). Any result can produce other results, the implementation of the way for the Effect generates Effect (inverse relationship is Effect generated By Effect). Both of these relationships are reversible. The reversibility of the relationship is important for the normalization of the Bayesian network and is applied to bidirectional reasoning. The ontology knowledge base constructed in this paper mainly comes from the field of hypertension, as shown in Fig.3.
The Proposed Bayesian Network Model

In the Bayesian network model, each node represents a domain variable, and the relationship between nodes represents the dependency between the node and the node. In this paper, the NETICA Application Programming Interface [11] (API) is used to create the Bayesian network. The hierarchical structure between the Bayesian network nodes can be automatically constructed by the program code, which is based on the hierarchical relationship between the entities in the medical ontology. All of the Bayesian network nodes are created in parallel in memory, and the parent-child relationship between nodes is created by using synchronization techniques between executing threads, so the child node will wait until all of its parent nodes are created. After the construction algorithm is completed, the conditional probability table will be filled with the default value (this operation is done automatically by NETICA). The structure of a conditional probability table depends on the number of parent nodes.

At the time of construction, there are three states: 1) for a root node without a parent node, the state of the conditional probability expression and the attribute defined by the node in the ontology are the same. 2) If the node has only one parent node, the conditional probability table is represented by a two-dimensional matrix. Where one dimension is the state of the node and the other is the state of the parent node. 3) For those nodes with more than one parent node, this paper uses the Noisy-Or algorithm [12] to reduce the number of probability parameters. Noisy-Or can greatly simplify the conditional probability table. And in practical applications it is easier to estimate the conditional probability between any two nodes than to estimate the joint conditional probability of multiple nodes for a node.

The Bayesian network can be constructed according to the following steps [13]:

Step 1 to determine the range of variables and variables in the range;
Step 2 to determine the relationship between variables in the field and describe the relationship as a graphical structure;
Step 3 The variables in the Bayesian are all derived from the entities defined in the body.

Definition 1 causality independence.

In the Bayesian network model, if the parent variables $X_1, X_2, \ldots, X_n$ can have a separate probability on the sub-variable $C$, it is said that the $n$ parent nodes are independent of each other. In the case of causality independence, the Bayesian network model is the Noisy-Or model, and the parent nodes of the nodes in the Noisy-Or model are causally independent. The conditional probability of occurrence of node $X_i$ in the Bayesian network shown in Fig. 4 is as follows.

$$P(X_i | X_1, X_2, X_3) = 1 - \prod_{i=1}^{n} (1 - P(X_i | X_i)).$$

Figure 3. Definition of ontology in the field of hypertension.

Figure 4. Simple Bayesian Network.
Most of the conditional Probability Table of Cardiovascular Disease Report [11]. Very few conditional probability tables are obtained by the cardiovascular physician under the guidance of statistics. Table 1 shows the impact of different risk factors on the risk of hypertension.

Table 1. Risk factors for hypertension in different risk factors.

<table>
<thead>
<tr>
<th>Risk factors</th>
<th>Risk factors for hypertension</th>
<th>Prevalence %</th>
</tr>
</thead>
<tbody>
<tr>
<td>Family Hypertension</td>
<td>No</td>
<td>18.22</td>
</tr>
<tr>
<td></td>
<td>Yes</td>
<td>30.38</td>
</tr>
<tr>
<td>Alcohol</td>
<td>No</td>
<td>24.04</td>
</tr>
<tr>
<td></td>
<td>Yes</td>
<td>28.83</td>
</tr>
<tr>
<td>Obesity</td>
<td>No</td>
<td>16.50</td>
</tr>
<tr>
<td></td>
<td>Yes</td>
<td>51.20</td>
</tr>
<tr>
<td>Triglyceride</td>
<td>Normal</td>
<td>20.69</td>
</tr>
<tr>
<td></td>
<td>High</td>
<td>37.20</td>
</tr>
<tr>
<td>Cholesterol</td>
<td>Normal</td>
<td>21.29</td>
</tr>
<tr>
<td></td>
<td>High</td>
<td>43.26</td>
</tr>
<tr>
<td>Smoking</td>
<td>No</td>
<td>22.54</td>
</tr>
<tr>
<td></td>
<td>Yes</td>
<td>26.32</td>
</tr>
</tbody>
</table>

Experiments and Analysis

NETICA provides a variety of programming language APIs that can manipulate the Bayesian network and provide a software product for visualizing the Bayesian network. NETICA’s visual reasoning experiment is as follows.

1) Generate Bayesian network hierarchy.

The input of the Bayesian network model algorithm automatically builds disease from the disease ontology, and the algorithm will also output a Bayesian network model that can be identified by the NETICA software product. However, this model cannot be used for reasoning, because the model contains only from the ontology mapping from the structural information.

2) Enter the digital information of the model - the conditional probability table.

For those nodes that do not have a parent node or only one parent node, their conditional probability table can be entered directly in the visual table. For those nodes with multiple parent nodes, the Noisy-Or algorithm is used to simplify the conditional probability table. The conditional probability will be calculated using equation (1). In the NETICA software, the conditional probability table of the Noisy-Or algorithm is expressed by the formula. In the Hypertension Bayesian network model, there are two nodes that need to use the Noisy-Or algorithm. The conditional probability formula for the node Obesity is shown in equation (2). Where the odds 0.30 for $p(\text{Obesity} \mid \text{Overnutrition}) = 0.30$, that is, Overnutrition (yes) is yes, Obesity is the probability of yes. Likewise, the number 0.27 can be similarly explained.

$$P(\text{Obesity} \mid \text{Overnutrition, not Enough Exercise}) = \text{Noisy - Or Dist} (\text{Overnutrition, 0.30, not Enough Exercise, 0.27})$$

The conditional probability formula for the node Hypertension as in Eq.(3).

$$\text{Hypertension, Obesity, triglyceride, Alcohol, Smoking, Family Hypertension)} = \text{Noisy Or Dist} (\text{Cholesterol, 0.4326, Obesity, 0.5120, high triglyceride, 0.3720, Alcohol, 0.2383, Smoking, 0.1632, Family Hypertension, 0.3038})$$

3) Enter patient evidence and diagnose.

Using the evidence shown in Table 2 to do the experimental test, the evidence from a diagnosis of patients with high blood pressure. From Table 2 enter to the node Obesity, Cholesterol, tri- glyceride all the evidence, based on the probability of suffering from hypertension is 65.2%.

The diagnostic results obtained from the input of these evidence are shown in Fig.5-7. At this point you can see the patient suffering from high blood pressure is 83.4%. Because of the stronger evidence
of Obesity, the evidence of the Overnutrition node and the not Enough Exercise node is lost. It can be seen from Fig. 6 that the probability of the Overnutrition node and the not Enough Exercise node is deduced by the evidence of the Obesity node. Finally, enter all the evidence in Table 2 and get the diagnostic conclusion shown in Fig. 7. At this point we can see that the probability of suffering from high blood pressure is 84.8%. In addition, the probability of high blood pressure in the patient may be further inferred from the probability of symptoms or other conditions caused by hypertension in the patient. In Fig. 5-7, the reason of the probability of these three nodes Cephalgia, Epistaxis, Dizziness is the default (50%) because we have not yet provided the conditional probability table for the relationship between hypertension and these nodes.

Figure 5. Diagnostic conclusion 1.

Figure 6. Diagnostic conclusion 2.

Figure 7. Diagnostic conclusion 3.

Table 2. System Test Data.

<table>
<thead>
<tr>
<th>Evidence</th>
<th>Probability</th>
<th>Evidence</th>
<th>Probability</th>
</tr>
</thead>
<tbody>
<tr>
<td>Over-nutrition</td>
<td>0.60</td>
<td>Triglyceride</td>
<td>1</td>
</tr>
<tr>
<td>Not Enough Exercise</td>
<td>0.55</td>
<td>Alcohol</td>
<td>1</td>
</tr>
<tr>
<td>Obesity</td>
<td>1</td>
<td>Smoking</td>
<td>0</td>
</tr>
<tr>
<td>Cholesterol</td>
<td>0</td>
<td>Family Hypertension</td>
<td>1</td>
</tr>
</tbody>
</table>

Conclusion

In this paper, ontology-driven Bayesian network model is proposed using the semantic ontology knowledge base, which automatically transforms the entities of the ontology into the Bayesian network model. First, hypertensive ontology is constructed to prove the validity of the model. In addition, the advantage of Bayesian network is applied for solving the uncertain and non-complete information. Finally, by mapping the hypertensive ontology in the Bayesian network model, the different probability of the condition is entered to obtain the probability that the patient is suffering from high blood pressure. The experimental results show that the ontology-driven Bayesian network model can make the correct judgment based on the data of hypertensive patients, and it is not difficult to see that the diagnostic accuracy depends on the patient's evidence. In addition, the accuracy of the diagnosis depends on the rationality of the Bayesian network structure and the accuracy of the conditional probability. The established model of hypertension can be used as a supplement to the existing method of diagnosis of hypertension, to assist the physician to make better diagnosis.
Reference


