Chapter 23
Clinical Decision Support Model of Heart Disease Diagnosis Based on Bayesian Networks and Case-Based Reasoning

Man Xu and Jiang Shen

Abstract To boost the accuracy of clinical decision support systems and degrade their misdiagnosis rates, a hybrid model was proposed with Bayesian networks (BN) and case-based reasoning (CBR). BN were constructed with the feature attributes and their casual relationships were learned. The similarities of feature attributes were measured with the case matching method, as well as the knowledge of their dependent relationships. Therefore, the accuracy of the diagnosis system was enriched through the dynamic retrieval method.

Keywords BN · CBR · Heart disease diagnosis

23.1 Introduction

Reasoning mechanism is one core component of clinical decision support systems (CDSS). Rule-based reasoning (RBR) is a popular inference technology in most of the existing expert systems (ES), as well as model-based reasoning (MBR). However, the relationship between information and knowledge is complex and fuzzy (Kong et al. 2008) in the area of healthcare, especially in the process of medical diagnosis. But the medical rules are more difficulty in acquiring and matching.

Cased-based reasoning (CBR) is the process of solving new problems based on the solutions of similar past problems. In CBR, processes like retrieval and matching are typically assumed to be broadly general cognitive processes.

M. Xu · J. Shen
TEDA College, Nankai University, Tianjin, China

M. Xu · J. Shen (✉)
College of Management and Economics, Tianjin University, Tianjin, China
e-mail: motoshen@163.com

E. Qi et al. (eds.), *The 19th International Conference on Industrial Engineering and Engineering Management*, DOI: 10.1007/978-3-642-38391-5_23,

Therefore, CBR is often used to solve complex problems with incomplete knowledge or difficulty in acquiring rules. The critical part of CBR is to find a suitable case retrieval method, which is determined to the efficiency and accuracy of the solution.

The process of case retrieval relies on the similarity measures for all the case attributes. The weights and similarities of features are given by physicians during medical diagnosis. Although the feature weights are taken into account with several algorithms, including feature evaluation (Shiu et al. 2001), introspective learning (Zhang and Yang 2001) and Neural Network (Zhang and Yang 1999), they omitted adaption strategies. Furthermore, Nearest Neighborhood (NN) method is adapted to case retrieval (Gu et al. 2003; Ling et al. 2006), but the solution cannot regenerate with database updating. The accuracy of NN decreases rapidly when dealing with incomplete data of medical problem, as well as its retrieval efficiency.

This paper built a hybrid model integrating CBR with Bayesian Networks (BN) for clinical decision supports (BN-CBR). Compared with conventional CBR, BN-CBR contributes to the optimization of the storage and retrieval process for medical cases, which improves the efficiency of CDSS. Moreover, BN-CBR aims to boost the accuracy of reasoning and to degrade the rate of misdiagnosis.

23.2 CDSS Based on the BN-CBR Hybrid Model

The diagnosis process of a clinician involves in synthesizing the information of a patient. This process also involves the circle of observation, diagnosis and treatment. From the perspective of clinic, the tasks of observation mainly consist of acquiring and mining the information of a patient completely, which reduces the uncertainty of the patient' condition to the greatest extent. With the available information, the clinician administers the treatment to the patient as well as his/her experience and medical knowledge. This diagnosis process can be regarded as a reasoning process, and its treatment is a solution for the medical problem. The treatment relies on the diagnosis and decision-making process, while its accuracy are determined by the belief and the adequacy of the acquired information during the period of observation (Ma et al. 2002). The framework of CDSS based on the BN-CBR model is demonstrated as Fig. 23.1.

As in Fig. 23.1, queries demonstrate the patients' condition (medical cases). The processor of CBR obtains the most similar case through the similarity measure. The conditional probabilities among the attributes of the cases are inferred through data training on Bayesian Networks (BN), which is an effective data mining tool. Therefore, the similarity function is formulated to measure the degree of similarity among cases. Medical knowledge and the information of patients are stored in the medical database, which is integrated into casebase after data screening and processing. The casebase as the main knowledge source for reasoning can be accumulated, refined and updated during the diagnosis reasoning episodes.

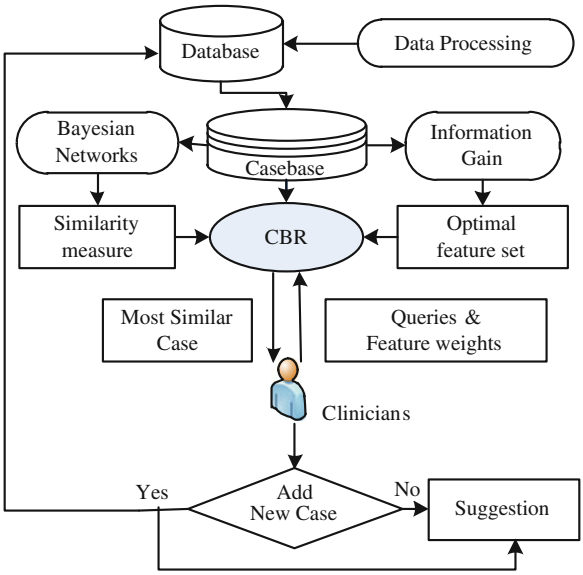

Fig. 23.1 The framework of CDSS based on the BN-CBR hybrid model

23.2.1 Knowledge Representation

Definition 23.1 A medical case is represented as

$$C_i = (Index, X_1, X_2, \ldots, X_m, S), \quad C_i \in CB, X_i \in X \tag{23.1}$$

where CB denotes the casebase, which is a finite set; X denotes the set of medical features; *Index* denotes the index of medical cases; S denotes the consequence or suggestion of the diagnosis.

Take the diagnosis of Coronary heart disease (CHD) for example. CB = (#, Age, Gender, Family history of CHD, chest pain type, blood pressure, cholesterol, class of CHD), where # is the index of cases. The class of CHD is binary, 0 indicating CHD while 1 indicating other health condition.

Definition 23.2 A medical case database is represented as $CB = [x_{ij}]_{m\times n}$, where x_{ij} denotes the jth attribute of the ith medical case, m denotes the number of features, and n denotes the size of casebase. CB consists of the specific knowledge of previous experiences or historical examples.

Definition 23.3 A medical query is represented as

$$MD = f(MP) = f(P, D, M, CB, K, M^+). \tag{23.2}$$

where MP denotes the medical problem; P denotes the patient; D denotes the possible states of the patient; M denotes the finite set of clinical symptoms; K denotes the knowledge set containing all the relationship among M, D and its causes of P; M^+ denotes the vital signs set of the patient, $M^+ \subset M$; DC denotes

the diagnose/conclusion for the patient, $DC \subset D$; $f(\cdot)$ denotes the reasoning function of the clinician's cogitation.

Clinical symptoms and vital signs are the prerequisite of the inference mechanisms in CDSSs. The basic idea is to drive the conclusion from the vital signs set M^+ based on the medical case database CB, the symptoms set M, the knowledge set K, the possible states D and the reasoning process f.

Definition 23.4 The reasoning process of CDSSs based on CBR is formulated as $f_{CBR}(C, T, Sim) = DC$, where T denotes the clinical query, and its dimensions represent the attributes; Sim denotes the similarity measure for two cases, and $Sim : (C_\mathrm{i}, T) \rightarrow (0, 1)$.

23.2.2 Feature Selection

Information gain (IG), also named as mutual information, is an important machine learning tool for weighting attributes. IG is the average value of information about the class with or without a feature. When the value of IG is larger, the information added by the feature is also much larger. Therefore, the values of IG for all the features are measured during the feature selection epoch. The feature with the largest value of IG belongs to the optimized feature set.

The idea of feature selection is to calculate the values of IG for all the attributes in the feature set, and to remove the features whose value of IG are less than a default threshold. The vital signs set M^+ and the symptoms set M is Initialized and the threshold is set as p. The feature f ($f \in M$) is traversed. Given a data sample, entropy

$$I(S_1, S_2, \ldots, S_m) = -\sum\nolimits_{i=1}^{m} P(C_i) \log_2 P(C_i) \tag{23.3}$$

where $P(C_i)$ is the prior probability for samples with C_i, $P(C_i) = S_i/S$; m denotes the number of classes, S_i the number of samples labeled by C_i, and S_i the total number of samples.

Definition 23.5 Assume that the feature f has v values, $\{f_1, f_2, \ldots, f_v\}$. The data sample is divided into v subsets by f, $\{S_1, S_2, \ldots, S_v\}$, where S_j contains the cases with feature value f_j. The term $S_{1j} + S_{2j} + \cdots + S_{mj}/S$ is the weight of the jth subset. S_{ij} is the number of samples with class C_i in S_j, then

$$I(S_{1j} + S_{2j} + \cdots + S_{mj}) = -\sum\nolimits_{i=1}^{m} P_{ij} \log_2 P_{ij} \tag{23.4}$$

where P_{ij} is the probability of samples with class C_i in S_j, $P_{ij} = S_{ij}/S_j$.

Information gain corresponding to *f is*

$$\begin{aligned} Gain(f) &= I(S_1, S_2, \ldots, S_m) - E(f), \\ E(f) &= \sum\nolimits_{j=1}^{v} \frac{S_{1j} + \cdots + S_{mj}}{S} I(S_{1j} + \cdots + S_{mj}). \end{aligned} \tag{23.5}$$

The feature is added into M^+ on the condition that $Gain(f)$ is larger than the threshold p.

23.2.3 Parameter Learning of BN

First, the conditional probability is calculated as $p(x_i, r_i|D_l, \theta^{(t)})$ for all the parameters in set D and the attribute x_i. Given set D, its likelihood is

$$l(\theta|D) = \sum_{l} \ln p(D|\theta) = \sum_{ijk} h(x_j^k, r_i^j) \ln \theta_{ijk}$$

where $h(x_j^k, r_i^j)$ denotes the assigned value in the database with $x_i = k$ and $r_i = j$. The maximum of the likelihood θ is

$$\theta_{ijk} = \frac{h(x_j^k, r_i^j)}{\sum h(x_j^k, r_i^j)} \tag{23.6}$$

Assuming the initial value $\theta^{(0)}$, the expectation of the current likelihood $\theta^{(t)}$ is

$$l(\theta|\theta^{(t)}) = \sum_{l} \sum \ln p(D_l, X_l|\theta) p(X_l|D_l, \theta^{(t)})$$

For any θ, if $L(\theta\Big|\theta^{(t+1)}) \geq L(\theta\Big|\theta^{(t)})$, then

$$L(\theta\Big|\theta^{(t)}) = \sum_{jlk} f\left(a_j^k, \pi(a_j)^l\right) \ln \theta_{jlk}$$

The next estimation of the likelihood is determined by finding its maximum likelihood (MLE),

$$\theta_{ijk}^{(t+1)} = \arg\max E[P(D|\theta)|D, \theta^{(t)}, S] \frac{f(x_j^k, r_i^j)}{\sum f(x_j^k, r_i^j)}$$

The conditional probability distribution is fixed through the above recursive algorithm, namely, the relation among the symptom features are formulated.

23.2.4 Case Retrieval

As mentioned above, case retrieval and matching is the core phase of CBR, in which the similarity is measured for the query with the historical cases. In the literature, the retrieval of CBR is to measure the difference among the features of cases, and Euclidean distance is an popular tool as the similarity measure function (Jain and Marling 2001).

$$Similarity(x, y) = -\sqrt{\sum_{i=1}^{m} (x_i - y_i)^2}$$

where x_i, y_i denote the observations of the ith feature ($f_i \in M^+$) corresponding to the case x and $y$respectively; m is the total number of features.

The problem is that this similarity assessment method omitted the cases during the updating episode, as a result of losing the relative information. Therefore, a new similarity measure function is proposed to solve the probability or BN-CBR.

Based on the BN of the features in M^+, the similarity of two medical cases x, y is

$$Similarity(x, y) = -\sqrt{\sum_{i=1}^{m} g(x_i, y_i)} \tag{23.7}$$

$$g(x_i, y_i) = \begin{cases} \|x_i, y_i\| & \text{when } x_i, y_i \text{ are numeric;} \\ 0 & \text{when } x_i \text{ is equivalent to } y_i, \\ & \text{and } x_i, y_i \text{ are symbolic;} \\ 1-p & \text{when } x_i, y_i \text{ are symbolic,} \\ & x_i = X, \text{ the observation of } y_i \text{ is missing;} \\ 1-q & \text{when } x_i, y_i \text{ are symbolic,} \\ & y_i = Y, \text{ the observation of } x_i \text{ is missing;} \\ 1 & \text{Others.} \end{cases}$$

where $p = P(y_i = X|y_1, \ldots, y_{i-1}, y_{i+1}, \ldots, y_m)$ and $q = P(x_i = Y|x_1, \ldots, x_{i-1}, x_{i+1}, \ldots, x_m)$. The probabilities are obtained with BN to reveal their dependent relationships.

The best case is initiated as $BESTCASE = 0$, as well as the best similarity. All the cases are retrieved with the formula (23.7) for the query y. If $Similarity(x', y) > BESTSIM$, then $BESTSIM = Similarity(x', y)$ and $BESTCASE = x'$. The final *BESTCASE* becomes the retrieved case.

23.3 Conclusion

This paper proposed the CDSS based on the BN-CBR hybrid model. BN were established with the feature attributes of heart disease, which improved the accuracy of diagnosis reasoning through solving the problem of missing data in

the database. Not only were the similarities of feature attributes measured, but also the case matching method integrated the knowledge of their dependent relationships.

Acknowledgments This work was supported by the National Natural Science Foundation of China (71171143), by Tianjin Research Program of Application Foundation and Advanced Technology (10JCYBJC07300), and Science and Technology Program of FOXCONN Group (120024001156).

References

Gu YS, Hua Q, Zhan Y et al (2003) Case-base maintenance based on representative selection for 1-NN algorithm. In: Presented at the international conference on machine learning and cybernetics, vol 4, pp 2421–2425

Jain AF, Marling CR (2001) Case-based tool for treatment of behavioral problems. In: Proceedings of the 33rd southeastern symposium on system theory, pp 337–341

Kong G, Xu DL, Yang JB (2008) Clinical decision support systems: a review on knowledge representation and inference under uncertainties. Int J Comput Intell Syst 1(2):159–167

Ling HF, Guo JY, Yan J (2006) Similarity algorithm of CBR technology applied to fault diagnosing field. J PLA Univ Sci Technol (Nat Sci Edn) 7(5):480–484 (in Chinese)

Ma XX, Huang XY, Huang M, Ni L (2002) Analysis of information flow based on entropy of fault diagnosis. J Chongqing Univ (Nat Sci Edn), 25(5):25–28 (in Chinese)

Shiu SCK, Yeung DS, Sun CH et al (2001) Transferring case knowledge to adaptation knowledge: an approach for case-base maintenance. Comput Intell 17(2):295–314

Zhang Z, Yang Q (1999) Dynamic refinement of feature weights using quantitative introspective learning. In: Proceeding of the international joint conference in artificial intelligence. Morgan Kaufmann, San Francisco, pp 228–233

Zhang Z, Yang Q (2001) Feature weight maintenance in case bases using introspective learning. J Intell Inf Syst 16(2):95–116