2014

ISSN : 0974 - 7435 Volume 10 Issue 21

BioTechnology

An Indian Journal

FULL PAPER

BTAIJ, 10(21), 2014 [13257-13261]

Medical data mining algorithm based on improved rough set theory and probabilistic neural network

Zhang Qiu-ju[1,2*], Li Jin-lin[1]
[1]School of Management and Economics, Beijing Institute of Technology, Beijing 100081, (CHINA)
[2]Institute of Systems Science andTechnology, Wuyi University, Jiangmen 529020, (CHINA)
E-mail: apollo5234@126.com

ABSTRACT

As medical information system is popularized in more hospitals. Since it can collect more information about patients' disease, it is feasible to use data mining technology to assist disease diagnosis. Based on rough set (RS) theory and PageRank algorithm, a new method was proposed to extract the key attributes of relevant attributes of diseases, and a probabilistic neural network (PNN) model was established for disease diagnosis. The results showed that the diagnostic accuracies of the model for patients with benign tumor and malignant tumor reached 100% and 95.24%, respectively, proving that the established model was effective and efficient in disease diagnosis.

KEYWORDS

Rough set; PageRank; Probabilistic neural network; Data mining; Medical management.

INTRODUCTION

In recent years, as medical information system is popularized in more hospitals, the system has collected lots of data about patients' disease. Deep analysis for the data using data mining technology can not only reveal the development law of disease, but also observe and summary all kinds of therapeutic regimens and therapeutic effects. These are significant for medical research, particularly the early diagnosis and prevention of disease.

At present, many studies been carried out by scholars in China and aboard concerning how to extract useful information from the above data sets[1-5]. Based on these studies, a disease diagnosis algorithm using data mining technology was proposed and verified by performing empirical analysis.

Among commonly used data mining algorithms, artificial neural network shows favorable advantages in pattern recognition, signal processing, knowledge engineering, expert system, optimum combination, robot control,*etc*... Probabilistic neural network (PNN) introduces Bayes classifying and decision-making theory in artificial neural network and separates decision space from multiple dimensional input spaces. Compared with traditional feedforward neural networks, PNN accomplishes learning fast in one time and exhibits high toleration for noise and distinct advantages in pattern classification. However, PNN calls for higher requirements for samples in classification and larger storage space. Therefore, to avoid dimension disaster and reduce storage space and processing time, dimension reduction for data is necessary in disease diagnosis by constructing PNN classification model.

As attribute reduction of data is non-deterministic polynomial (NP) complete problem, its solution takes a long time. Aiming at this, PageRank algorithm was referred to reduce the attributes by ranking attribute significances and eliminating redundant attributes.

DISEASE DIAGNOSIS MODEL

Data attribute reduction

There are numerous complex data in medical data mining. Generally, to ensure the accuracy of diagnosis, conditions (condition attributes) of patients have to be learned as many as possible. However, not all of the conditions (condition attributes) are useful for the diagnosis. In PNN model construction, the input of these numerous independent variables evidently influences operation efficiency. Therefore, it is necessary to perform knowledge reduction, which is one of the core ideas of rough set (RS) theory as well.

In RS theory, U is universe and R is its equivalence relation. They compose an approximation space $\mathrm{K} = (\mathrm{U}, \mathrm{R})$. If $(\mathrm{x}, \mathrm{y}) \in \mathrm{U}$ and $(\mathrm{x}, \mathrm{y}) \in \mathrm{R}$, then x and y are undistinguishable in K, and R is an indiscernibility relation. If attribute set $\mathrm{P} \subseteq \mathrm{R}, \mathrm{P} \neq \phi$, then $\cap \mathrm{P}$ is an indiscernibility relation in P, and denoted as $\mathrm{ind}\,(\mathrm{P})$. If the attribute $\mathrm{r} \in \mathrm{P}$ and $\mathrm{ind}\,(\mathrm{P}) = \mathrm{ind}\,(\mathrm{P} - \{\mathrm{r}\})$, then r is omissible in P; otherwise r is necessary in P. If all $\mathrm{r} \in \mathrm{P}$ are necessary in P, P is independent. If $\mathrm{Q} \subseteq \mathrm{R}$, Q is independent, and meets $\mathrm{ind}(\mathrm{Q}) = \mathrm{ind}\,(\mathrm{R})$, then Q is a reduction of R. If Q is a reduction of R, then the dependability of decision attribute (D for short) onQ is equal to the dependability of decision attribute on R: $\gamma_C(D) = \frac{|Pos_Q(D)|}{|U|} = \gamma_R(D) = \frac{|Pos_R(D)|}{|U|}$.

To efficiently reduce attributes using RS theory, vote principle of PageRank is adopted to sort the significances of different condition attributes. PageRank evaluates the significance of a webpage depends on its vote number, which is determined by the significance of webpages that it links to. Each hyperlink to the webpage is a vote. Therefore, two important factors are involved in the algorithm, hyperlink number and weights of voting webpages.

When rank the significances of condition attributes using PageRank algorithm, to eliminate redundant attributes as many as possible, the hyperlink numbers of two condition attributes have to be inversely proportional to their correlation coefficients. If they are totally linearly correlated, they can be replaced by and not vote to each other. Here $v_{ij} = v_{ji} = 1 - r_{ij}$ represents the number that attribute c_i and c_j vote to each other (r_{ij} is the correlation coefficient between c_i and c_j), and the correlation coefficient r_{iD} between condition attribute and decision attribute demonstrates the weight of attribute c_i voting to other attributes. Therefore, the relative significance $sig(c_j) = \sum_{i=1}^{m} r_{iD} \times v_{ij}$, and m is the number of condition attributes.

But there is a disadvantage in the above method for calculating relative significances of attributes. That is, when suppose condition attribute c_j is highly correlated with other condition attributes, then according to the above method, these attributes cast little number of votes to c_j. As a result, the relative significance of c_j reduces and c_j is likely to be eliminated initially as redundant attribute by ranking the attribute significances. However, the attribute, which is highly correlated with other attributes, is possible the key attribute if it is also highly correlated with decision attribute. Therefore, the formula for calculating the relative significance of attribute is modified. Additionally, on the basis of RS theory, the following attribution reduction algorithm is designed:

Step 1: Disease diagnosis result (decision attribute) is D, which is divided into difference categories and represented using different integers. For example, a diagnosis result can be divided into two categories, diagnosed with certain disease and without the disease, which are illustrated using 1 and 0, respectively.

Step 2: The multiple correlation coefficients of all the attributes in condition attribute set C and the correlation coefficients of each attributes with decision attribute D are calculated.

Step 3: The maximum correlation coefficients of each condition attribute$\mathrm{c_i}$ with others are selected and demonstrated as $\mathrm{Max_i = max\{r_{ij}, i \neq j, j = 1,2,\dots,m\}}$. Suppose that $\mathrm{Max_i = r_{ik}}$, and$\mathrm{temp_i = \{c_i, c_k\}}$, then the relative significance $\mathrm{sig\ (temp_i) = \sum_{j=1}^{m} r_{jD} \times v_{ij} + \sum_{j=1}^{m} r_{jD} \times v_{kj}}$, and m is the number of condition attributes.

Step 4: If$\mathrm{c_i \in temp_j}$, then $\mathrm{a_j = 1}$, otherwise, $\mathrm{a_j = 0}$. The relative significance of $\mathrm{c_i}$ is denoted as $\mathrm{sig\ (c_i) = 1/2 \times \sum_{j=1}^{m} a_j \times sig\ (temp_j)}$.

Step 5: Condition attribute set C is discretized. As RS can merely process discrete attributes, if an attribute value of the original data is continuous, its value range can be divided into several intervals. Each interval has a unique code ranging within 1,2,3,…,k, among which k is the number of the intervals. The attribute values of each sample are illustrated using their corresponding interval codes.

Step 6: The condition attributes are ranked in terms of their relative significances and renumbered. The condition attribute with minimum significance is denoted as $\mathrm{c'_1}$, followed by $\mathrm{c'_2}$, which merely larger than$\mathrm{c'_1}$. Similarly, all the condition attributes are renumbered.

Step 7: Let $\mathrm{i = 1, C_{reduce} = C}$.

Step 8:$\mathrm{c'_i}$ is removed from condition attribute set C. Then the dependency degrees $\mathrm{\gamma_C(D)}$ and $\mathrm{\gamma_{C-\{c'_i\}}(D)}$ of D for C and $\mathrm{C - \{c'_i\}}$ are calculated. If they are equal, then $\mathrm{C\_reduce = C\_reduce - \{c'_i\}}$, and $\mathrm{i = i + 1}$.

Step 9: If $\mathrm{i \leq m}$, Step 8 is repeated; otherwise, $\mathrm{C\_reduce}$ is output.

PNN diagnosis model

PNN, which was put forward by Specht in 1990, is a kind of artificial neural network subjecting to statistical principle. Its network structure includes input layer, pattern layer, summation layer, and output layer. In input layer, all the neurons are input singly and output singly with linear transfer functions. The layer expresses the input signals distributively. Pattern layer and input layer is connected by connection weight ω_{ij}. The transfer functions of neurons in pattern layer are non-linear operators and the number of neurons equals to input samples. Summation layer linearly summarizes the inputs transferred from pattern layer. In the layer, the number of neurons is that of categories which are intended to be divided. Output layer has judgment function and its neurons are output as discrete values, which represent the categories of input patterns.

According to PNN theory, the algorithm below is designed:

Step 1: Normalization processing is performed for the data to obtain data set W.

Step 2: The data set W is divided into training set and test set, and $\mathrm{Train \cup Test = w, Train \cap Test = \emptyset}$.

Step 3: PNN model is trained using data set P.

Step 4: Input data in P, T are simulated using trained PNN model and the simulation results are output.

Step 5: Simulation results are compared with actual disease diagnosis results to obtain diagnostic accuracy.

EMPIRICAL ANALYSIS

Relevant data about breast neoplasms diagnosis are downloaded from UCI data set. It contains 9 attribute indexes and 1 diagnosis result index. The 9 attribute indexes are Clump Thickness, Uniformity of Cell Size, Uniformity of Cell Shape, Marginal Adhesion, Single Epithelial Cell Size, Bare Nuclei, Bland Chromatin, Normal Nucleoli and Mitoses.

By applying the algorithm designed, a reduced attribute set consisting of 5 attribute indexes is obtained: Clump Thickness, Uniformity of Cell Shape, Marginal Adhesion, Bare Nuclei, and Mitoses. This reduction of condition attributes set is denoted as C_reduct. UsingMATLAB to do the calculation, it costs less than 2 minutes to finish the whole process, and $\gamma_{C\_reduct}(D) = \gamma_C(D) = 0.9224$.

The reduced attribute set is divided into training set and test set, among which the former contains 500 samples and the later contains 183 samples. PNN is constructed using the inner function $\mathrm{newpnn\ (P, T, Spread)}$ of MATLAB, among which P is input vector, T is objective vector, Spread parameter is the expanding coefficient of radial basis function. Tests for the applied data set are conducted repeatedly and the constructed PNN is trained when $\mathrm{Spread = 0.1}$. Finally, the diagnosis results are obtained by diagnosing diseases using the trained PNN model.

The training set contains 303 samples diagnosed with benign tumor and 197 with malignant tumor.

Because the correct recognition of malignant tumor is more important than the correct identification of benign tumor, in order to guarantee the correct recognition rate of malignant tumor, we can adjust the penalty coefficient of misjudging malignant tumors.

When the penalty coefficient is 1, by training PNN model using data in the training set, 303 benign samples and 196 malignant ones are recognized. The overall classification accuracy of the training set reaches 99.8%; then the data in the test set are classified using the model as well. It recognizes 141 benign samples and 39 malignant ones from 141 patients diagnosed with benign tumor and 42 ones with malignant tumor. The overall classification accuracy of the test set is 98.36%.

When the penalty coefficient goes to 2, and other parameters keep unchanged, a new PNN model is trained and applied to classification of tumor. The classification results are as shown in TABLE 2. The new model identifies 1 more malignant tumor sample both in test set and in training set, While the correct identification rate of benign tumor is still high.

The classification results indicate that the model exhibits high classification accuracy for tumor. Considering overall results, both the classification accuracies of training set and test set exceed 98%. It implies that the high classification accuracy of the model is not because of the learning process of modeling data. Therefore the model can be popularized in actual application.

TABLE 1 :Accuracy of PNN classification model using C_reduct (Penalty coefficient=1)

	The diagnosis results	**The classification results of PNN model**	**Accuracy (%)**
Training set	benign tumor (303)	benign tumor (303)	100
	malignant tumor (197)	malignant tumor (196)	99.49
Test set	benign tumor (141)	benign tumor (141)	100
	malignant tumor (42)	malignant tumor (39)	92.86

TABLE 2: Accuracy of PNN classification model using C_reduct(Penalty coefficient=2)

	The diagnosis results	**The classification results of PNN model**	**Accuracy (%)**
Training set	benign tumor (303)	benign tumor (300)	99.01
	malignant tumor (197)	malignant tumor (197)	100
Test set	benign tumor (141)	benign tumor (141)	100
	malignant tumor (42)	malignant tumor (40)	95.24

If we use some classic method, for example, genetic algorithm to obtain the reduction of C, and MATLAB to do the calculation, itwill take 15-20 minutes, while the parameter values of population size and generations are both 100. The reduction of C is not unique. In 10 experiments, the reduction appears most frequently consists of Clump Thickness, Uniformity of Cell Shape, Marginal Adhesion, Single Epithelial Cell Size, and Bare Nuclei. This reduction of condition attributes set is denoted as C_reduct', and $\gamma_{C_{reduct}}{}'(D) = \gamma_C(D) = 0.9224$.
The penalty coefficient of misjudging malignant tumors remains 2, and other parameters keep unchanged, the classification results are as shown in TABLE 3.

TABLE 3: Accuracy of PNN classification model using C_reduct'(Penalty coefficient=2)

	The diagnosis results	**The classification results of PNN model**	**Accuracy (%)**
Training set	benign tumor (303)	benign tumor (303)	100
	malignant tumor (197)	malignant tumor (197)	100
Test set	benign tumor (141)	benign tumor (140)	99.29
	malignant tumor (42)	malignant tumor (40)	95.24

The classification accuracy in TABLE 3 has no big difference from TABLE 2. But it costs more time to obtain C_reduct'. Take together, the algorithm designed in this paper is not only effective, but also more efficient.

CONCLUSIONS

PNN model for diagnosing breast neoplasms is constructed based on RS theory and PNN and verified by empirical analysis for 683 actual samples of a hospital acquired from UCI data set. The results indicate that:

1) By applying attribute reduction method of RS theory and PageRank, a mathematical model was established and 5 key attributes were obtained by reducing the data of breast neoplasms. Comparing with classic reduction algorithm of rough set based on GA, reduction algorithm of rough set based on PageRank was proved to be more efficient.

2) A PNN model was established based on the reduced data and then utilized for diagnoses of patients. The results showed that the model constructed is highly accurate, with practical application prospect.

REFERENCES

[1] J.Soni, U.Ansari, D.Sharma, et al.; Predictive data mining for medical diagnosis: An overview of heart disease prediction, International Journal of Computer Applications, **17(8)**, 43-48 **(2011)**.

[2] Hsiang-Yang Chen, Chao-Hua Chuang, Yao-Jung Yang, et al.; Exploring the risk factors of preterm birth using data mining, Expert Systems with Applications, **38(5)**, 5384-5387 **(2011)**.

[3] S.B.Patel, P.K.Yadav, D.P.Shukla; Predict the Diagnosis of Heart Disease Patients Using Classification Mining Techniques, IOSR Journal of Agriculture and Veterinary Science, **4(2)**, 61-64 **(2013)**.

[4] H.Kaur, S.K.Wasan; Empirical Study on Applications of Data Mining Techniques in Healthcare, Journal of Computer Science, **2(2)**, 194-200 **(2006)**.

[5] S.C.Huang, E.C.Chang, H.H.Wu; A case study of applying data mining techniques in an outfitter's customer value analysis, Expert Systems with Applications, **36(3)**, 5909–5915 **(2009)**.

[6] J.C.Hsieh, C.C.Tai, M.S.Su, Y.H.Lin; Identification of partial discharge location using probabilistic neural networks and the fuzzy c-means clustering approach, Electric Power ComponSyst, **42(1)**, 60–9 **(2014)**.

[7] F.Modaresi, S.Araghinejad; A Comparative Assessment of Support Vector Machines, Probabilistic Neural Networks, and K-Nearest Neighbor Algorithms for Water Quality Classification, Water Resources Management, **28(12)**, 4095-4111 **(2014)**.

[8] O.Antropov, Y.Rauste, H.Astola, et al.; Land Cover and Soil Type Mapping From SpacebornePolSAR Data at L-Band With Probabilistic Neural Network, IEEE Transactions on Geoscience and Remote Sensing, **52(9)**, 5256 - 5270 **(2014)**.

[9] GaoXue-xing, Sun Hua-gang, Hou Bao-lin; A Neural Network Learning Method Using Samples with Different Confidence Levels, Journal of Electronics & Information Technology, **36(6)**, 1307-1311 **(2014)**.

[10] Ming-Shou Su, Chung-Chu Chia, Chien-Yi Chen, et al.; Classification of partial discharge events in GILBS using probabilistic neural networks and the fuzzy c-means clustering approach, Electrical Power and Energy Systems, **61**, 173–179 **(2014)**.