Original Paper

Establishing Assistant Diagnosis Models of Solitary Pulmonary Nodules Based on Intelligent Algorithms

Zhijun Zhao^a Jingtao Chen^a Xiaoxiang Yin^a Huayong Song^a Xinchun Wang^a
Jing Wang^b

^aDepartment of Oncology, The First Hospital Affiliated to Henan University, Kaifeng, ^bDepartment of Imaging, The First Affiliated Hospital of Zhengzhou University, Zhengzhou, China

Key Words

Solitary pulmonary nodule • Intelligent algorithm • Tumor markers • CT imaging features • Diagnosis model

Abstract

Aims: This study aimed to establish an auxiliary diagnosis model for solitary pulmonary nodules (SPNs) and evaluate its test efficacy. **Methods:** Three hundred thirty-two pathologically diagnosed SPN patients (186 malignant, 146 benign) were collected as subjects. The serum levels of 8 types of markers and 9 computed tomography (CT) imaging features of each patient were treated as independent variables. The corresponding pathological classification results (fungal inflammation, tuberculosis and tuberculoma, inflammatory pseudotumor, tumor middle differentiation, cancer) of quantized patients were treated as dependent variables. A 17-to-1 mathematical auxiliary SPN diagnosis model was established using a back propagation (BP) algorithm and a support vector machine (SVM) algorithm. A 40-case test set was used to estimate the effect. **Results:** Two different auxiliary SPN diagnosis models were successfully established. The diagnostic accuracy, sensitivity and specificity of the BP algorithm diagnosis model were 60%, 68% and 46.7%, respectively, and those of the SVM algorithm model were 80%, 85.7% and 66.7%, respectively. **Conclusion:** The accuracy, sensitivity and specificity of the SVM diagnostic model were relatively high, indicating that the model has important reference value for determining the degree of SPN differentiation and is suitable for the auxiliary diagnosis of benign and malignant SPN.

Copyright © 2015 S. Karger AG, Basel

Introduction

Solitary pulmonary nodules (SPNs) refer to quasi-circular solid lesions with diameters less than or equal to 3 cm inside the lung. They are not accompanied by lymphadenectasis, atelectasis, pneumonia or other diseases. The differential diagnosis for SPN is implemented for injuries caused by infection, inflammation, benign tumors, malignant tumors, blood vessel (vascular) abnormalities and congenital deformities, among others. Diagnosis and prognosis of this disease, to a large extent, rely on the properties of the SPN itself. Therefore, determining whether an SPN is benign or malignant is crucial in clinical practice. High resolution, low noise ratios, stable image quality and high definition of solitary nodules are important factors in finding the nodules. These lesions are very common, and it is often difficult to make a definite diagnosis. On the one hand, the enlarged nodule is relatively easy to diagnose, but on the other, this is contradictory to the early detection, diagnosis and treatment of tumors. The preferred method for the detection and diagnosis of pulmonary nodules, at present, is morphological evaluation with chest CT, but it is still very difficult to determine the diagnosis by morphology only. In addition to morphology evaluation, current clinical diagnosis also mainly refers to the nodule's doubling time and hemodynamic features, the metabolic characteristics of pulmonary nodule ^{18}F -FDG and the pathology results obtained from needle biopsy and wedge resection of the lung by thoracoscope [1]. At the same time, computer-aided diagnosis models such as artificial neural networks, support vector machines (SVMs) and linear regression models have increasingly been applied in medical diagnosis [2-4]. However, there is still no mature theory with high efficiency and accuracy that can evaluate the diagnosis of SPN, which limits the disease's timely diagnosis and treatment. Thus, using comparative analysis of the simple diagnostic models established by the BP and SVM algorithms to differentiate SPN degree, this study can provide reliable tools for SPN diagnosis, which further improves clinical diagnostic accuracy and is of great significance in developing timely and effective therapeutic schedules.

Materials and Methods

The study subjects

The subjects were 332 SPN patients who received treatment at the First Affiliated Hospital of Zhengzhou University from 2010 to 2013. There were 186 cases with malignant nodules, 106 males and 80 females, who ranged in age from 20 to 71 years old with an average age of 45.7 ± 24.3 years; there were also 146 cases with benign nodules, 82 males and 64 females, aged from 23 to 74 years, with an average age of 50 ± 26.6 years. All nodules were confirmed by pathology. Because the definitions of solid versus semi-solid solitary pulmonary nodules are unclear, and the distinction is not widely used in clinical practice [5], the random samples in this study were not limited to those with solitary pulmonary nodules.

Collecting and processing iconographic and serologic characteristic data

After the considerations of expert opinions and previous studies, we selected 8 serum markers as indexes: carcinoembryonic antigen (CEA), neuron-specific enolase (NSE), cytokeratin 19 fragments (CYFRA21-1), carbohydrate antigen 125 (CA125), cancer antigen 199 (CA199), cancer antigen 724 (CA724), squamous cell carcinoma (SCC), and carbohydrate antigen 153 (CA153). We also selected 9 SPN image feature parameters: size, location, boundary, spicule sign, lobulation, pleural indentation, vessel convergence sign, lymphadenectasis and cavity sign. All 17 indexes were treated as diagnostic parameters of SPN. The detailed data extraction is shown in Table 1.

The artificial neural network transfer function requires that the input parameters be distributed in the range [0, 1]. Therefore, parameters that did not meet the input requirements were treated using normalization processing by the linear function transfer. The formula is $Y = (X - \text{Min Value}) / (\text{Max Value} - \text{Min Value})$, where X and Y are the values of the input parameters before and after transferring, respectively.

Using the SPN differentiation degree and clinical data, this study divided the pathology results of SPN into ten types, namely, cancer, poorly differentiated tumor, low differentiated tumor, middle differentiated

Table 1. Data extraction of the serology and iconographic characteristics

Parameters	Indexes	Data extraction
Serum markers	CEA	—
	NSE	—
	CYFRA21-1	—
	CA125	—
	CA199	—
	CA724	—
	SCC	—
	CA153	—
Iconographic characteristics	SPN size	The average values of the anteroposterior diameter, transverse diameter and vertical diameter of the mediastinal window were obtained.
	SPN position	The value of the superior lobe was 10, the middle lobe was 5 and the inferior lobe was 0, among which, the hilus pulmonis was classified into the middle lobe and the lingual lobe into the superior lobe.
	Boundary	The edges of the lung nodules and pulmonary parenchyma were compared, and the values were assigned between the values of 0 and 10 from clear to dim.
	Spicule sign	The short, thin unbranched radial line was bounded to 5 mm. 0 was assigned as no sentus, and 1~5 were assigned as long sentus from long to short. 6~10 were assigned as short sentus from long to short.
	Lobulation sign	0 was assigned to no lobulation. If the number of outline evaginations between every two adjacent incisurae was 1, 2, 3, or 4, then a value of 2, 4, 6, or 8 was assigned, respectively. If the number was more than 4, a value of 9 or 10 was assigned.
	Pleural indentation	Presence of this syndrome was assigned 10, otherwise 0.
	Vessel convergence sign	Absence of this sign was assigned 0. Its presence was assigned 1~10 according to the degree.
	Lymphadenectasis	Absence was assigned 0, unobvious 5 and exact lymphadenectasis 10.
	Cavity	Absence was assigned 0, and presence was assigned 10.

tumor, inflammatory, high differentiated tumor, well differentiated tumor, inflammatory pseudotumor, hamartoma and hemangioma, tuberculosis and tuberculoma and fungal inflammation. Because the original study sample of poorly, low, and high differentiated tumors, well differentiated tumors, and hamartoma and hemangioma was low (generally, fewer than 10 cases), and to avoid inaccurate training results caused by limited time spent training the diagnostic model, these five types were not included in the output pathological type. Ultimately, the pathology results of this study were summarized into five categories, namely, fungal inflammation, tuberculosis and tuberculoma, inflammatory pseudotumor, middle tumor differentiation, cancer and were assigned as follows: 0.1, 0.2, 0.4, 0.7, and 1.

Research methods

Error back propagation (BP) algorithm. The commonly used methods in researching network models include BP algorithms, feedback network models and self-organizing network models, among others.

BP networks have self-learning and generalization ability; they can learn and store large numbers of input-output mode mappings without previously revealing the mathematical equations that describe the mapping relationships. In addition, BP is the most widely used form of neural network. The support vector machine (SVM), proposed by Corinna and Vapnik, can solve small samples, nonlinearity, high dimensionality and other challenges. In this study, the BP and SVM algorithms were used for programming.

Neural networks based on the BP algorithm are typical feedforward layered systems. The main thoughts of this algorithm were divided into two phases in this study. In the first phase (positive signal propagation), the input signal went through the input layer and the hidden layer and was processed step by step. In addition, the actual output value of each node was calculated. In the second phase (backward propagation error correction), if the expected output value was not obtained in the output layer, then the error between the actual and expected output was calculated layer by layer recursively. In accordance with the error, the weights were corrected until the expected output was attained [6]. Figure 1 shows the structure of the BP network.

Two hundred ninety-two samples were selected randomly from the 332 samples as the training samples and constituted the training set, and the remaining 40 samples were used as the test set. Then,

Fig. 1. Structural representation of the BP network.

Fig. 2. The SVM system.

MATLAB R2012a (The MathWorks, Inc., Natick, MA, USA) was used to establish a three-layer BP network model with a single hidden layer. The following parameters were set in the training process. (1) Excitation functions. Two functions were needed for the network in training, the transfer function from the input layer to the hidden layer and that from the hidden layer to the output layer. In this training, the applied transfer function from the input layer to the hidden layer was the tansig function, and the applied reference transfer function from the hidden layer to the output layer was either the logsig or the purelin function. (2) Training function. Different training methods were used for different training functions. The training functions applied in this model were trainrp and trainoss. (3) Input layer nodes. There were 17 input parameters in this model, and thus the number of input layer nodes was 17. (4) The number of hidden layers. Previous experiments demonstrated that a hidden layer could realize any classification decision, and thus, one hidden layer was used in this study [6]. (5) Number of hidden layer nodes. At present, no universal theory of determining the number of hidden layer nodes has been developed. Generally, it is believed that the number of hidden layer nodes appeared between the numbers of input and output layers. During the training, preliminary experiments were conducted by trial and error to determine the number, and it was finally indicated that when the node number of hidden layers was set at 8 or 10, better results were achieved. (6) One (1) was treated as the node number of the network output layer. (7) Ten thousand (10 000) was regarded as the threshold value of training times. (8) The accuracy was 0.001. In training the BP network, different parameter combinations achieved different effects. In this experiment, cross-combinations were applied to the functions of the input and output layers, the training function, hidden layer nodes and other parameters. Additionally, multiple combination approaches were adopted to train the samples, and those with the best combination effects were retained.

Support vector machine (SVM). The SVM is a Vapnik-Chervenenkis (VC) dimension theory established on the basis of statistical learning theory and a common learning method based on risk minimization principles. Based on the error rate (i.e., the generalization error rate) for the learning machine's test data, the machine took the sum of the training error rate and an item that relied on that VC dimension as the bound. Under the decomposable mode, the SVM value for the former item was 0, and the second was minimized. The SVM system is shown in Figure 2.

Fig. 3. The error fitting curve.

Fig. 4. The training results.

The SPN auxiliary diagnosis model was built with the SVM algorithm. All cases were divided into five categories: fungal inflammation, tuberculosis and tuberculoma, inflammatory pseudotumor, tumor middle differentiation, or cancer. Then, each category was further divided into two groups. After the data were recombined, one group was treated as the training set and the other as the test set. Five classifiers (SVM0, SVM1, SVM2, SVM3 and SVM4) were constructed using the LIBSVM3.17 software proposed by Chih-Jen Lin et al., and the current popular sequential minimal optimization algorithm was adopted. Seventeen-dimensional pathological feature and imaging feature vectors were taken as the input vectors [7].

This study used a parallel grid search method to select the SVM parameters. The grid search algorithm gives M and N values to the penalty parameter C and kernel parameter γ , respectively, to obtain $M \times N$ (C, γ) combinations. Training different SVMs and then estimating the promotion recognition rate gives a combination with the highest rates in the $M \times N$ (C, γ) combinations, and this is identified as the optimal parameter. After verification, the ultimately determined optimal parameter (C, γ) was (12, 1).

Evaluating the different SPN diagnosis models. After the prediction model was established, the model was used in the differential diagnosis of the remaining 40 cases to determine its validity. Commonly used indicators of evaluating diagnosis models include the following:

$$\text{Accuracy rating} = (\text{TP} + \text{TN}) / (\text{TP} + \text{TN} + \text{FP} + \text{FN}) \times 100\%$$

$$\text{Sensitivity} = \text{TP} / (\text{TP} + \text{FN}) \times 100\%$$

$\text{Specificity} = \text{TN} / (\text{TN} + \text{FP}) \times 100\%$, where TP is the true positives; TN is the true negatives; FP is the false positives; and FN is the false negatives.

Results

The BP algorithm model

The BP network was trained by applying various combination approaches. After multiple tests and screening, the final selected parameter combinations were as follows. The excitation function in the input layer was tansig. The excitation function in the output layer was purelin, and the training function was trainrp. The node number of the hidden layers was 10, the error threshold value was 0.001, and the threshold value of the training times was 10000. The error fitting curve is shown in Figure 3. As observed in the figure, when the training time reached 5715, the model accuracy had reached 0.001 and the training ended.

The training result is shown in Figure 4. In this figure, the vertical coordinates are the actual output of the BP network, and the abscissa represents the desired target outputs. When the actual output was equal to the target output, the sample points are located on the diagonal line. This figure displays that most sample points after the training were located around the diagonal line, indicating better training results.

Fig. 5. The BP test curve.

Fig. 6. The SVM simulation results.

Table 2. Diagnostic results of comparing the two models

Pathology results	BP algorithm				SVM			
	TP	TN	FP	FN	TP	TN	FP	FN
Cancer	6	2	2	2	8	2	1	1
Middle differentiated tumor	3	1	3	3	9	3	1	1
Inflammatory pseudotumor	4	2	1	1	3	1	1	1
tuberculosis and tuberculoma	1	1	1	1	1	1	1	1
Fungal inflammation	3	1	1	1	3	1	0	0
total	17	7	8	8	24	8	4	4
sensibility	68%				85.7%			
accuracy	60%				80%			
specificity	46.7%				66.7%			

The 40 test samples were entered into the established model for simulation; then, a fit comparison of the simulation and actual results was performed, and those results are presented in Figure 5. The figure indicates that the fit between the simulation and the actual pathological results was low. Some samples had a larger deviation degree, which showed that the effect was not ideal.

The SVM model

The 40 test sample cases were entered into the SVM model for training and simulation. The simulation results are displayed in Figure 6, where the hollow circle is the target SVM outputs and the solid circle is the actual simulation output. It was found that the fit between the SVM model simulation and actual pathologic results was greater and the effect was better.

Diagnosis effect comparison of two models

We identified the diagnostic accuracy of the two mathematic diagnostic models built by the BP and SVM algorithms (Table 2). It was found that the diagnostic accuracy using the BP algorithm was 60%, whereas that using SVM was 80%. The diagnostic sensibility and specificity of SVM were both higher than those of the BP algorithm.

Discussion

Because of the lack of clinical specificity and the diverse morphological image characteristics of SPN, in addition to the fact that sputum and bronchoscopy examination results cannot be confirmed, qualitative SPN diagnosis has been both a focus of and a difficulty for clinicians, and it is often directly related to the choice of treatment and prognosis, which is also a worldwide challenge.

To solve the problem of qualitative diagnosis with limited accuracy of image diagnostics, this study established two diagnostic models using the BP algorithm and SVM.

The BP algorithm is a nonlinear pattern classifier. By training the historical data, the BP neural network can memorize the samples it has learned into the connection weights of each neuron through learning. As long as the convergence of the network training is within the required error accuracy, all of the training sample outputs are correct. At the same time, the BP network has a certain generalization ability and good classification ability for learning samples. However, the BP algorithm needs a large number of network training parameters that are set artificially, and it is prone to generating locally optimal solutions [8]. In addition, the number of layers and nodes in the network's hidden layer network has no theoretical guidance and is generally determined based on experience or through trial and error. Therefore, there is often a great deal of network redundancy, which to a certain extent also increases the network's learning burden.

SVM exhibits many unique advantages in solving problems such as small samples, nonlinearity and high dimensional pattern recognition. To a large extent, it has overcome difficulties in dimensionality and over-fitting as well [9]. As a more advanced approach in machine learning, SVM has been increasingly used as a classifier in SPN segmentation and extraction [10-12]. However, SVM has been used in only the diagnosis of the nodular degree of differentiation and was used by Campadelli [13] to detect SPNs in only plain chest film and the predetermined quantification of identified prostate tumors in MRS images.

In this study, the nine selected imaging features are the most commonly used iconographic characteristics for reference when differentiating SPN, the main diagnostic bases for radiologists and also the main long-standing bases of computer-aided diagnosis [14, 15]. At the same time, the serum tumor marker has the merits of being noninvasive, easy to check, and fast, in addition to having broad application [16]; as such, it has also been widely used for the diagnosis of pulmonary diseases. In this study, based on the nine iconography features and eight serum markers, two diagnostic models established by the BP algorithm and SVM were comparatively analyzed. The results showed that the diagnostic accuracy of the SVM model reached 80%, and its sensibility and specificity were 85.7% and 66.75%, respectively, both higher than the findings for the diagnostic model based on the BP algorithm. Thus, SVM is more suitable for clinical needs.

In general studies, mathematic diagnostic models have been established to identify benign and malignant SPNs; for example, the logistic regression analysis by Tian R manifested that spicule sign, gender, smoking and age were all independent factors that affected SPN properties, and a relevant mathematic model [$\text{Logit}(P) = -8.722 + 2.448(\text{gender}) + 2.023(\text{smoking}) + 0.851(\text{age}) + 1.057(\text{diameter}) + 2.432(\text{spiculation}) + 1.502(\text{FDG uptake})$] was established to predict the SPN malignance probability [17]. Li Y and Wang J also used a mathematic model to predict benign and malignant SPNs. Their logistic regression analysis results also indicated that 6 clinical symptoms—spicule sign, SPN diameter and verges—were independent risk factors for malignant SPN. Moreover, the sensitivity, specificity,

positive predictive and negative predictive values of the prediction model established by the 6 risk factors were 94.5%, 70.0%, 87.8% and 84.8%, respectively [18]. By detecting eight scanning features—the average gray value, gray value variance, area, diameter, circulation, rectangularity, centroidal moment and centroidal angle of the pulmonary nodule—and extracting four relative features that were relevant to the verge and calcification, Cao L et al [19] treated the features as a 12-D feature vector and performed a qualitative analysis of pulmonary nodules with few training samples using SVM. In this study, the model output results were divided into five levels, fungal inflammation, tuberculosis and tuberculoma, inflammatory pseudotumor, middle differentiated tumor, or cancer, which also meant that the pulmonary diseases were classified into five levels from benign to malignant. Thus, the qualitative diagnosis of SPN was realized, which is rare in existing research in China. This study resolves the challenge of accurately diagnosing pulmonary nodules as benign or malignant, and the results are beneficial for pathologic diagnosis results and patient treatment. Our model has very important significance in improving the early diagnosis and treatment of lung cancer.

Conclusion

The support vector machine, as an assistant diagnostic tool, belongs to the category of computer-aided diagnosis and cannot fully replace clinicians' diagnoses [20-22], although its diagnostic value has been universally recognized. The SPN assistant diagnostic model based on SVM in this study has considerable significance in diagnosing varying degrees of lung disease differentiation, and it can be used as a powerful tool for lung disease diagnosis.

Acknowledgements

We thank the Medical Engineering Technology & Data Mining Institute of Zhengzhou University for their technical instructions and writing assistance and the First Hospital Affiliated to Henan University for their material support.

Disclosure Statement

The authors declare that there are no conflicts of interest regarding the publication of this paper.

References

- 1 Xu SC, Cheng T: Solitary pulmonary nodules: detection and characterization with medical imaging and guideline for further diagnosis and treatment. *Int J Med Radiol* 2011;34:141-145.
- 2 Lin DT, Yan CR, Chen WT: Autonomous detection of pulmonary nodules on CT images with a neural network-based fuzzy system. *Comput Med Imaging Graph* 2005;29:447-458.
- 3 Frank EH, Grodzinsky AJ, Koob TJ, Eyre DR: Streaming potentials: asensitive index of enzymatic degradation in articuLar cartilage. *J Orthop Res* 1987;5:497-508.
- 4 Buckwalter JA, Roughley PJ, Rosenberg LC: Age-related changes in cartilage proteoglycans: quantitative electron microscopic studies. *Microsc Res Tech* 1994;28:398-408.
- 5 Naidich DP, Bankier AA, MacMahon H, Schaefer-Prokop CM, Pistolesi M, Goo JM, Macchiarini P, Crapo JD, Herold CJ, Austin JH, Travis WD: Recommendations for the Management of Subsolid Pulmonary Nodules Detected at CT: A Statement from the Fleischner Society. *Radiology* 2013;266:304-317.
- 6 Liu D: The application of BP based on genetic algorithm in medical diagnosis. Jilin, Jilin University, 2007.

- 7 Zhou Y, Zhang XS, Zhang HX, Wu J, Liu L, Ma JW, Liu WY: Research on the solitary pulmonary nodule diagnostic model based on CT images. *J Pract Radiol* 2011;27:26-34.
- 8 Burden F, Winkler D: Bayesian regularization of neural networks. *Methods Mol Bio* 2008;458:25-44.
- 9 Ding SF, Qi BJ, Tan HY: An overview on theory and algorithm of support vector machines. *J U Electron Sci Technol China* 2011;40:2-10.
- 10 Sun T, Wang J, Li X, Lv P, Liu F, Luo Y, Gao Q, Zhu H, Guo X: Comparative evaluation of support vector machines for computer aided diagnosis of lung cancer in CT based on a multi-dimensional data set. *Comput Meth Prog Bio* 2013;111:519-524.
- 11 Wang Q, Kang W, Wu C, Wang B: Computer-aided detection of lung nodules by SVM based on 3D matrix patterns. *Clin Imaging* 2013;37:62-69.
- 12 Böröczky L, Zhao L, Lee KP: Feature subset selection for improving the performance of false positive reduction in lung nodule CAD. *IEEE Trans Inf Technol Biomed* 2006;10:504-511.
- 13 Campadelli P, Casiraghi E, Artioli D: A fully automated method for lung nodule detection from postero-anterior chest radiographs. *Ieee Trans Med Imaging* 2006;25:1588-1603.
- 14 Lesperance LM, Gray ML, Burstein D: Determination of fixed charge density in cartilage using nuclear magnetic resonance. *J Orthop Res* 1992;10:1-13.
- 15 Farndale RW, Butler DJ, Barret AJ: Improved quantitation and discrimination of sulfated glycosaminoglycan by use of dimethylmethylene blue. *Biochim Biophys Acta* 1986;883:173-177.
- 16 Alberg JA, Ford JG, Samet JM: Epidemiology of lung cancer. *Chest* 2007;132:29-55.
- 17 Tian R, Su MG, Tian Y, Li FL, Kuang AR: Development of a predicting model to estimate the probability of malignancy of solitary pulmonary nodules. *Sichuan Da Xue Xue Bao Yi Xue Ban* 2012;43:404-408.
- 18 Li Y, Wang J: A mathematical model for predicting malignancy of solitary pulmonary nodules. *World J Surg* 2012;36:830-835.
- 19 Cao L, Li WJ, Feng QJ: Automatic detection and diagnosis of lung nodules on CT images based on LDA and SVM. *Nan Fang Yi Ke Da Xue Xue Bao* 2011;31:324-328.
- 20 Harders SW: LUCIS: lung cancer imaging studies. *Dan Med J* 2012;59:B4542.
- 21 Choromanska A, Macura KJ: Evaluation of solitary pulmonary nodule detected during computed tomography examination. *Pol J Radiol* 2012;77:22-34.
- 22 Ellis MC, Hessman CJ, Weerasinghe R, Schipper PH, Vetto JT: Comparision of pulmonary nodule detection rates between preoperative CT imaging and intraoperative lung palpation. *Am J Surg* 2011;201:619-622.