Association for Information Systems
AIS Electronic Library (AISeL)

PACIS 2016 Proceedings

Pacific Asia Conference on Information Systems (PACIS)

Summer 6-27-2016

A NEW METHOD FOR PREDICTING EARLY-STAGE LUNG NODULES BASED ON PSO-SVM HYBRID ALGORITHM

Shan Li
Nanjing University of Aeronautics and Astronautics, lishan@nuaa.edu.cn

Ying Yu
Nanjing University of Aeronautics and Astronautics, yuy0711@163.com

Haibin Chen
The first affiliated hospital with Nanjing Medical University,, wayne0403@163.com

A NEW METHOD FOR PREDICTING EARLY-STAGE LUNG NODULES BASED ON PSO-SVM HYBRID ALGORITHM

Shan Li, College of Economics and Management, Nanjing University of Aeronautics and Astronautics, Nanjing, China, lishan@nuaa.edu.cn

Ying Yu, College of Economics and Management, Nanjing University of Aeronautics and Astronautics, Nanjing, China,yuy0711@163.com

Haibin Chen. The first affiliated hospital with Nanjing Medical University, Nanjing, China, wayne0403@163.com

Abstract

The aim of this article was to use the Support Vector Machine (SVM) to predict the benign and malignant solitary pulmonary nodules (SPNs) in early-stage lung cancer in order to lessen the patient's pain and save the money. Fifty and one patient records were collected .Each record consisted of four clinical characteristics and nine morphological characteristics. The SVM classifier was built by radial basis kernel function. The penalty factor C and kernel parameter σ were optimized by comparing particle swarm optimization (PSO), grid search algorithm (GSA) and genetic algorithm (GA)and then employed to diagnose the SPNs. By comparison with a Logistic regression (LR) model, the overall results of our calculation demonstrated that the area under the receiver operator characteristic (ROC) curve for the model (0.913 ± 0.051, $p<0.05$) was higher than the LR model. The accuracy, sensitivity and specificity in the model were 90.7%, 89.3% and 93.3% respectively. It is represented that the PSO-SVM model can be used in predicting the early-stage lung nodules.

Keywords: Lung cancer; Support vector machine; Prediction; Diagnosis; Particle swarm optimization; Grid search; Genetic algorithm

1 INTRODUCTION

Lung cancer is one of the most common visceral malignant tumor in the world. Each year about 1.1 million people died of lung cancer (Yano et al. 2010). The high morbidity and mortality of lung cancer have seriously threated human health. Comparing to the other cancers, the biological characteristics of lung cancer are very complex. Lung cancer, in its early stages, is asymptomatic or just with mild symptoms, difficult to be detected. The cancer cell easily transfers to other parts of the body in a short time , if you don't take any treatment. Moreover, late lung cancer is difficult to be cured. Therefore, it is the major and key way for physicians to detect and treat in the early growth stages of tumor, which can improve the survival rate of lung cancer patients. Recent researches have shown that if lung cancer in the early was detected and treated timely, the patient survival rate may rise from 14% to 49% (Chapman et al. 2008). Computer Tomography (CT) scans are widely used to diagnose with pulmonary diseases. These diseases usually behave as Solitary Pulmonary Nodules (SPNs) in imaging. Thus, the detection and recognition of SPNs are the best way to diagnose the pulmonary diseases. But in pulmonary CT images, lung cancer always confuses with some benign lesions such as pulmonary tuberculosis or inflammatory pseudotumor, because these diseases are also characterized by nodules in CT images. Confronting such complicated problem, clinical experience and judgment may not be reproducible or reliable, whereas a quantitative model might have advantages in accuracy and reproducibility, will not be uninfluenced by personal judgment, and can provide outcome exchange ability.

Many previous studies have shown the usefulness of computer-aided prediction models based on clinical data for early-stage lung cancer detection. Two widely cited logistic regression (LR) models were proposed by Swensen et al. (1999) and Gould et al. (2007) with their area under the curve (AUC) of the receiver operating characteristic (ROC) of 0.83 ± 0.02 and 0.79 ± 0.05, respectively. Due to the maturation of machine learning (ML) theory and technologies, decision trees (DTs) (Zinovev et al. 2012) and artificial neural networks (ANNs) (Shiraishi et al. 2011; Chen et al. 2012; Kuruvilla &Gunavathi 2014) are then used to classify a nodule as either malignant or benign. However, most clinical data are incomplete which gathered quite difficultly. They are usually small samples which may not apply to ANNs model and DTs model. Fortunately, support vector machine (SVM) exhibits many unique advantages in solving problems such as small samples, nonlinearity and high dimensional pattern recognition (Zhao et al. 2015). It is for these reasons that SVM manifests such powerful diagnostic capabilities in the prediction of some deadly diseases such as breast cancer (Sabatier et al. 2011; Schrauder et al. 2012), coronary heart disease (Giri et al. 2013; Kruppa et al. 2014).

Studies using SVM model for properties of SPNs have been published outside of China, Sousa et al. (2010) explored six stages to extract texture for automatic detection of lung nodules in CT images using support vector machines. Kim et al. (2009) used 11 shape features and 13 textural features, with support vector machine and Bayesian classifiers, to improve performance of differentiating obstructive lung diseases, based on high-resolution computerized tomography (HRCT) images. In the work proposed by Keshani et al. (2013), a system for lung nodule detection, segmentation and recognition using CT was presented. The lung area was segmented using active contours, then a masking technique was used to transfer non-isolated nodules into isolated ones. Nodules were detected using a SVM with 2D stochastic and 3D anatomical features. In the paper studied by Madero et al.(2015), the nodules were characterized by the computation of the texture features obtained from the gray level co-occurrence matrix (GLCM) in the wavelet domain and were classified using a SVM with radial basis function in order to classify CT images into two categories: with cancerous lung nodules and without lung nodules.

Nonetheless, most of these models were only built on morphological characteristics data without including clinical characteristics data. In the past decade, several researchers (Erasmus et al. 2000; Swensen et al.1999) whose studies have been published have indicated that clinical characteristics

such as smoking and age have become risk factors of lung disease. Thus, in this paper, we both collected morphological characteristics data and clinical characteristics data to create a SVM model for judging characteristics of SPNs. Comparing with other classification algorithms, the results displayed that the prediction of the SVM model is better, and achieved higher classification accuracy.

2 METHODS

2.1 Clinical data

The subjects were 94 SPN patients who received treatment at the first affiliated hospital with Nanjing Medical University. Among these 94 SPNs, 51 SPN patients were randomly selected as train set named group A. There were 30 cases with malignant nodules, 13 males and 17 females, who ranged in age from 45 to 79 years old with an average age of 64 ± 10 years; there were also 21 cases with benign nodules, 12 males and 9 females, aged from 36 to 69 years, with an average age of 55.5 ± 9.3 years. The rest of the clinical data were collected as the test set named group B. All nodules were confirmed by pathology. Clinical data including clinical characteristics data (age of the patient, gender, smoking history, history of previous cancer diagnosis) and morphological characteristics data (diameter of tumor, spiculation, lobulation, calcification, pleural retraction sign, border, cavity, ground-glass opacity, and pointed sign) were collected.

2.2 Support vector machine model

Support vector machine is a supervised learning method based on statistical learning theory and the structural risk minimization principle (Zinovev et al. 2011). Using the training data, SVM implicitly maps the original input space into a high dimensional feature space. Subsequently, in the feature space the optimal hyper plane is determined by maximizing the margins of class boundaries. The training points that are closest to the optimal hyper plane are called support vectors. Once the decision surface is obtained, it can be used for classifying new data. The aim of the SVM classification is to find an optimal separating hyper plane that can distinguish the two classes from the mentioned set of training data (Pradhan 2013).

Given training vectors $x_i \in R^n, i = 1,..., l$, in two classes, and an indicator vector $y_i \in R^l$ such that $y_i \in \{-1,1\}$.

The hyper plane of SVM classification is described as:

$$\omega \cdot x_i + b = 0 \tag{1}$$

Where ω is a coefficient vector that determines the orientation of the hyper plane in the feature space, b is the offset of the hyper plane from the origin.

The primal optimization problem is defined as:

$$\begin{aligned} &\min \ \frac{1}{2}\|\omega\|^2 + C\sum_{i=1}^{l}\xi_i \\ &\text{s.t. } y_i \cdot ((\omega \cdot \varphi(x_i)) + b) \geq 1 - \xi_i, i = 1,2,..., l, \\ &\xi_i \geq 0, i = 1,2,..., l \end{aligned} \tag{2}$$

Where $\varphi(x_i)$ maps x_i into a higher-dimensional space, $C > 0$ is the regularization parameter, ξ_i is the positive slack variables. Due to the possible high dimensionality of the vector variable ω , usually we solve the following dual problem using Lagrangian multipliers:

$$\min \sum_{i=1}^{n} \alpha_i - \frac{1}{2} \sum_{i=1}^{n} \sum_{j=1}^{n} \alpha_i \alpha_j y_i y_j (x_i x_j)$$

$$\text{s.t.} \sum_{i=1}^{n} \alpha_i y_j = 0$$

$$0 \le \alpha_i \le C \tag{3}$$

The decision function, which will be used for the classification of new data, can then be written as:

$$g(x) = sign(\sum_{i=1}^{n} y_i \alpha_i x_i + b) \tag{4}$$

In cases when it is impossible to find the separating hyper plane using the linear kernel function, the original input data may be transferred into a high dimension features space through some nonlinear kernel functions. The classification decision function is then written as:

$$g(x) = sign(\sum_{i=1}^{n} y_i \alpha_i K(x_i, x_j) + b) \tag{5}$$

Where $K(x_i, x_j)$ is the kernel function.

The performance of the SVM model depends on the choice of the kernel parameters. In literature, there are several kernel types. However, four kinds of them are often used: linear kernel, polynomial kernel, RBF kernel and sigmoid kernel as a last one. The RBF Kernel is by far one of the most powerful kernels. In many studies and cases (especially in nonlinear problems), RBF performs the best prediction results (Pradhan 2013). For this reason, in this paper, RBF kernel was employed for kernel.

2.3 Parameters optimization

For the RBF kernel, the regularization parameter(C) and the kernel width (σ) are the two parameters that need to be optimized. Grid search is often used for parameter determination. The search process consists of varying parameters by a fixed step size through a wide range of values and then evaluating the performance of each combination. Because of its computational complexity, grid search is only suitable for the optimization when there are very few parameters (Friedrichs & Igel 2005). With the development of heuristic optimization methods, certain optimization techniques such as the genetic algorithm (GA) (Huang & Wang 2006) and particle swarm optimization (PSO) (Huang & Dun 2008) have been adopted in parameter optimization for SVM.

Genetic algorithm is a programming technique that mimics biological evolution as a problem-solving strategy. Given a specific problem to solve, the input to the GA is a set of potential solutions to that problem, encoded in some fashion, and a metric called a fitness function that allows each candidate to be quantitatively evaluated. These candidates may be solutions already known to work, with the aim of the GA being to improve them, but more often they are generated at random.

Particle swarm optimization is an emerging population-based meta-heuristic that simulates social behavior such as birds flocking to a promising position to achieve precise objectives in a multidimensional space. Like evolutionary algorithms, PSO performs searches using a population (called swarm) of individuals (called particles) that are updated from iteration to iteration. To discover the optimal solution, each particle changes its searching direction according to two factors, its own best previous experience (pbest) and the best experience of all other members (gbest).

This paper genetic algorithm, grid search algorithm and particle swarm optimization algorithm as the SVM parameters optimization algorithms were compared in order to improve the classification

accuracy. Tools named LibSVM are used which were developed by Dr.Lin (Chih-Jen Lin) from Taiwan University.

2.4 Logistic regression model

The LR model is a kind of generalized linear regression model that is widely used to estimate the probability of a dichotomous outcome event being related to a set of predictors. In the previous study (Li & Wang 2012), a LR model with six independent variables (two continuous and four categorical) corresponding to the six selected features was developed as $p = e^x/(1+e^x)$, x = -4.496 + (0.07*age) + (0.676*diameter) +(0.736*spiculation) + (1.267*family history of cancer) - (1.615*calcification) - (1.408*border), where e is the natural logarithm, and the value for the last four elements, i.e., family cancer history, calcification, spiculation, and border, equals 1 if the element exists, and 0 otherwise. A p value of 0.463 was ultimately selected as a cut-off point and p values>0.463 should be considered malignant disease and p<0.463 should be considered benign.

3 RESULTS

3.1 Data analysis

The data analyzed in this study included 21 cases of benign disease (42%) and 30 cases of malignant disease (58%) in group A. We employed four clinical characteristics (age of the patient, gender, smoking history, history of previous cancer diagnosis) and nine morphological characteristics (diameter of tumor, spiculation, lobulation, calcification, pleural retraction sign, border, cavity, ground-glass opacity, pointed sign) of lung nodules on CT images. Thereinto, two features included patient age and the nodule's diameters are quantitative characteristics, while other features are qualitative characteristics. We used SPSS13.0 software (2004, IBM, Armonk, NY) for statistical analysis. The analysis results are shown in Table 1.

characteristic	Malignant(n=30)		Benign(n=21)		P-value[1]
	n	%	n	%	
Gender					0.4
Female	17	56.7	9	42.9	
Male	13	43.3	12	57.1	
Smoking history					**<0.05**
Yes	18	60	4	23.5	
No	12	40	17	76.5	
History of previous cancer diagnosis					0.634
Yes	1	4.8	3	10	
No	20	95.2	27	90	
Spiculation					**<0.05**
Yes	7	33.3	23	76.7	
No	14	66.7	7	23.3	
Lobulation					**<0.05**
Yes	7	33.3	25	83.3	
No	14	66.7	5	16.7	
Ground-glass opacity					0.193
Yes	3	14.3	10	33.3	

[1] using Fish's Exact Test

No	18 85.7	20 66.7	
Border Smooth Not smooth	 1 4.7 20 95.3	 3 3.3 27 96.7	0.634
Pleural retraction sign Yes No	 4 19 17 81	 15 50 15 50	**<0.05**
Cavity Yes No	 2 9.5 19 90.5	 8 26.7 22 73.3	0.167
Pointed sign Yes No	 1 4.7 20 95.3	 3 3.3 27 96.7	0.634
Calcification Yes No	 1 4.7 20 95.3	 4 13.3 26 86.7	0.391
	mean±STD	mean±STD	P-value[2]
Age	64 ± 10	55.5 ± 9.3	**<0.05**
Diameter	2.22±0.68	1.39±0.99	**<0.05**

Table 1. Statistical results of the features between the patients with malignant and with benign nodules in group A

3.2 Comparison of SVM parameters optimization algorithms

3.2.1 SVM parameters optimization based on PSO (PSO-SVM)

Based on experience, we set the regularization parameter $C \in [0.1,100]$ and the kernel width $\sigma \in [0.1,1000]$. PSO parameters were set as follows:

The initial population is set to 20; C_1, C_2 are acceleration constants that satisfy C_1=1.5, C_2=1.7; W=1, where W represents inertia weight; R is denoted maximum number of iterations and R= 200; fitness function value was equaled the classification accuracy of 5-Fold cross-validation. The GA fitness curve was shown in the Figure 1.

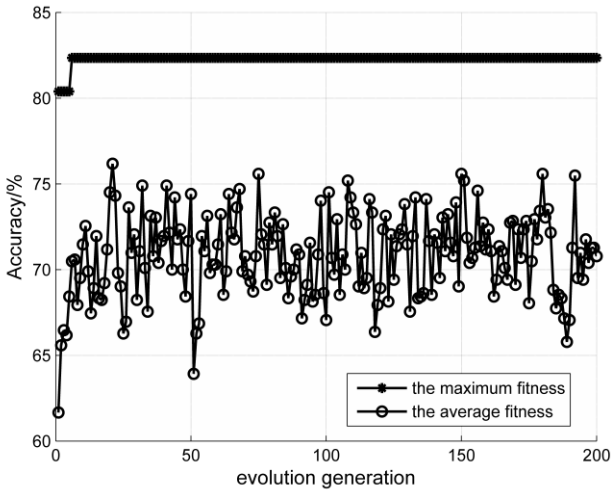

Figure 1. Accuracy curve in PSO-SVM with Group A dataset

[2] using independent-samples T test

As demonstrated in the Figure1, after the genetic algorithm ran to 25 generations, the average fitness has been increased to maximum value then tended to be stable, and the maximum fitness value increased until the algorithm ran to 20 generations. At this point, the output of the SVM optimal parameter C = 88.21, σ= 0.01.

3.2.2 SVM parameters optimization based on GSA (GSA-SVM)

Based on experience, we set the regularization parameter $C \in [2^{-8}, 2^{8}]$ and the kernel width $\sigma \in [2^{-8}, 2^{8}]$. GSA parameters were set as follows: The step length C_p and σ_p are 0.8; fitness function value was equaled the classification accuracy of 5-Fold cross-validation. The GSA fitness curve was shown in the Figure 2.

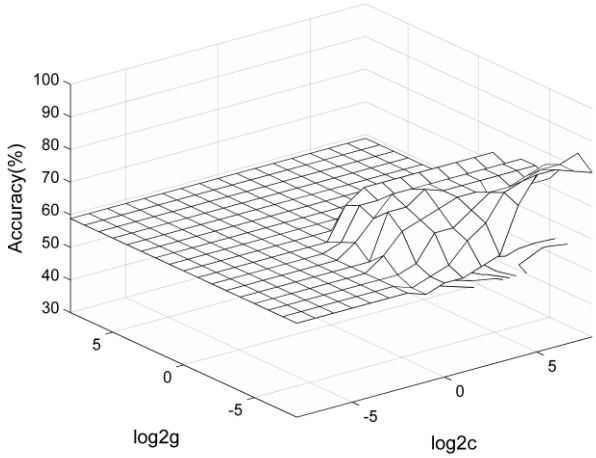

Figure 2. Accuracy curve in GSA-SVM with Group A dataset

Figure 2 was shown that classification worked best when $\log 2c \in [-10,10]$, $\log 2g \in [-10,10]$. At this point, the output of the SVM optimal parameter C = 32, σ= 0.0039.

3.2.3 SVM parameters optimization based on GA (GA-SVM)

Based on experience, we set the regularization parameter $C \in [0,100]$ and the kernel width $\sigma \in [0,1000]$. GA parameters were set as follows:

The initial population is set to 20; R is denoted maximum number of iterations and R= 200; fitness function value was equaled the classification accuracy of 5-Fold cross-validation. The GA fitness curve was shown in the Figure 3.

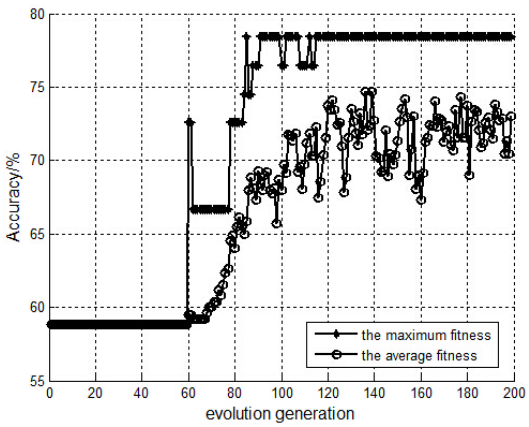

Figure 3. Accuracy curve in GA-SVM with Group A dataset

As demonstrated in the Figure 3, after the genetic algorithm ran to 120 generations, the average fitness has been increased to maximum value then tended to be stable, and the maximum fitness value increased until the algorithm ran to 115 generations. At this point, the output of the SVM optimal parameter C = 3.32, σ= 0.087.

3.2.4 Evaluating the different SVM models

Three common indicators were used to compare the results obtained:

$$accuracy(Acc) = \frac{TP + TN}{TP + TN + FP + FN} \tag{6}$$

$$sensitivity(Sen) = \frac{TP}{TP + FN} \tag{7}$$

$$Specificity(Spe) = \frac{TN}{TN + FP} \tag{8}$$

$$Matthews\ correlation\ coefficient\ (Mcc) = \frac{TP * TN - FP * FN}{\sqrt{(TP + FP)(TP + FN)(TN + FP)(TN + FN)}} \tag{9}$$

Where TP is the true positives; TN is the true negatives; FP is the false positives; and FN is the false negatives.

				Group B dataset							
Methods	C	σ	Acc[3]	TP	FN	TN	FP	acc/%	sen/%	spe/%	Mcc
N-SVM	1	0.5	-	26	2	11	4	86.1	92.9	73.3	0.687
PSO-SVM	88.21	0.01	82.4	25	3	14	1	90.7	89.3	93.3	0.805
GSA-SVM	32	0.0039	84.3	26	2	11	4	86.1	92.9	73.3	0.687
GA-SVM	3.32	0.087	78.4	25	3	12	3	86.0	89.3	80.0	0.693

Table 2. Results of the different SVM models

The comparison result is shown in Table 2. In this Table, N-SVM represents the SVM model without parameters optimization (default). Compared to N-SVM and GSA-SVM, we found that there was no significant difference. However, when used PSO-SVM, the accuracy and specificity are better than N-SVM. In addition, the accuracy and the Matthews correlation coefficient exhibit a significant improvement among PSO-SVM, GSA-SVM and GA-SVM, This result indicates that the PSO-SVM approach can determine the parameter values without lowering SVM classification accuracy in this study. Thus, the PSO-SVM model was suitable for clinical needs.

3.3 Diagnosis effect comparison of PSO-RBF-SVM model and LR model

After the prediction model was established, the model was used in the differential diagnosis of the remaining 43 cases to determine its validity. Then we identified the diagnostic accuracy of the two mathematic diagnostic models built by the LR and SVM algorithms (Table 3). It was found that the

[3] represent the accuracy of train set (Group A)

diagnostic accuracy using the LR algorithm was 72.1%, whereas that using SVM was 90.7%. Although the diagnostic sensitivity of LR was higher than of the SVM algorithm, it was shown that the specificity of LR was only 26.7%, which indicated the misdiagnosis rate was high.

Methods	TP	FN	TN	FP	acc/%	sen/%	spe/%	Mcc
PSO-RBF-SVM	25	3	14	1	90.7	89.3	93.3	0.805
LR	27	1	4	11	72.1	96.4	26.7	0.343

Table 3. Comparison of results between PSO-RBF-SVM and LR classifiers

In order to obtain information on how accurately the SVM and LR distinguish subjects with different outcomes, the receiver operating characteristic (ROC) curve was computed. The ROC curve is a popular and powerful tool to assess discrimination for binary outcomes. The curve is created by plotting the true positive rate against the false positive rate at various threshold settings (Mulshine & Smith 2002). The ROC curve obtained for the SVM and LR was shown in Figure 4 and Figure 5. The area under curve (AUC) value of SVM and LR models were 0.913±0.051 and 0.765±0.074, respectively.

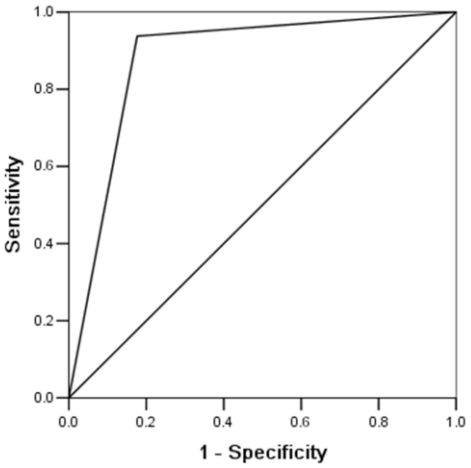

Figure 4. Receiver operator characteristic (ROC) curve generated using SVM model

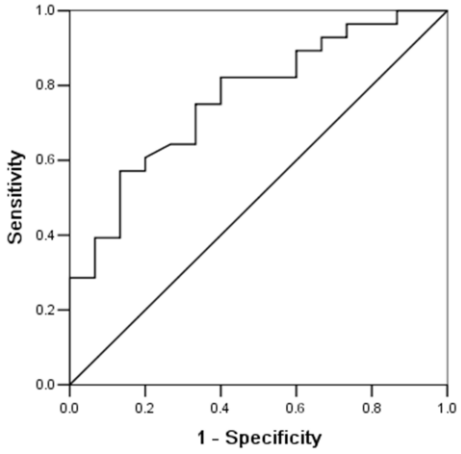

Figure 5. Receiver operator characteristic (ROC) curve generated using LR model

4 DISCUSSION

Because of the lack of clinical specificity and the diverse morphological image characteristics of SPNs, in addition to the fact that sputum and bronchoscopy examination results cannot be confirmed, qualitative SPNs diagnosis has been both a focus of and a difficulty for physicians, 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 clinical data, this study established a diagnostic model named PSO-RBF-SVM. Up until now, such SVM model has been studied by Zhao et al. However, to surpass the previous work, our study gives the more comprehensive data collection of both clinical and morphological characteristics while Zhao et al. only used morphological characteristics. In the past decade, several researchers (Erasmus et al. 2000; Swensen et al.1999) whose studies have been published have indicated that clinical characteristics such as smoking and age have become risk factors of lung disease. These two factors were also shown significant differences (both $p<0.05$) in Table 1. There is no doubt that more comprehensive the data collected, the more

reliable the model will be. Thus our model is more accurate (Acc for models of us and Zhao et al. are 90.7% and 80%, respectively) because it is based on comprehensive and systematic data collection.

In general studies, mathematic diagnostic models have been established to identify benign and malignant SPNs. This study a LR model developed by Li & Wang compared with our PSO-RBF-SVM model. Results were shown that PSO-RBF-SVM outperformed LR model in terms of accuracy and specificity, although PSO-RBF-SVM were inferior to LR models in terms of sensitivity (Table 3).The AUC has been recommended as a better discrimination measure, since it may be interpreted as the average sensitivity across all possible specificities and the average specificity across all possible sensitivities, and it has a higher convergence than the accuracy rate (Glas et al. 2003). Therefore, PSO-RBF-SVM had a higher discriminant performance than LR model in terms of AUC (AUCs for PSO-RBF-SVM and LR models were 0.913 ± 0.051 and 0.765 ± 0.074 ($p < 0.05$), respectively).

In this study, neither the LR models nor the PSO-RBF-SVM were tested by external validation. Data were analyzed retrospectively and the results were based on a series of information obtained on a relatively small group of cases in terms of diagnosis and clinical characteristics. Further work will collect more clinical data to obtain the more optimal SVM classification for physicians to diagnose the pulmonary diseases more accurately.

5 CONCLUSION

The SNPs assistant diagnostic model based on SVM in this study has considerable significance in diagnosing the properties of SPNs and it can be used as a powerful tool for lung disease diagnosis. However, the support vector machine remains as an assistant diagnostic tool in that it cannot be used as a substitute for a pathologic diagnosis, although its diagnostic value has been universally recognized. Therefore, physicians need to seriously consider all SPNs by combining with diagnostic results of the model.

6 ACKNOWLEDGEMENTS

Supported by the Jiangsu Province Natural Science Foundation (Serial Number: BK2012385); the Research Fund for the Doctoral Program of Higher Education of China (Serial Number: 20123218120034); the Fundamental Research Funds for the Central Universities (Serial Number: NS2013083).

References

Yano T, Yamazaki K, Maruyama R, Tokunaga S, Shoji F, Higashi H, Lung Oncology Group in K. 2010. Feasibility study of postoperative adjuvant chemotherapy with S-1 (tegaful, gimeracil, oteracil potassium) for non-small cell lung cancer-LOGIK 0601 study. Lung Cancer, 67(2):184-187.

Chapman CJ, Murray A, McElveen JE, Sahin U, Luxemburger U, Tureci O, Robertson JF. 2008. Autoantibodies in lung cancer: possibilities for early detection and subsequent cure. Thorax, 63(3): 228-233.

Swensen SJ, Silverstein MD, Edell ES, Trastek VF, Aughenbaugh GL, Ilstrup DM, Schleck CD. 1999. Solitary Pulmonary Nodules: Clinical Prediction Model Versus Physicians. Mayo Clinic Proceedings, 74(4):319-329.

Gould MK, Ananth L, Barnett PG, Veterans A. 2007. A clinical model to estimate the pretest probability of lung cancer in patients with solitary pulmonary nodules. Chest, 131(2):383-388.

Zinovev D, Duo Y, Raicu DS, Furst J, Armato SG. 2012. Consensus versus disagreement in imaging research: a case study using the LIDC database. J Digit Imaging, 25(3):423-436.

Shiraishi J, Li Q, Appelbaum D, Doi K. 2011. Computer-Aided Diagnosis and Artificial Intelligence in Clinical Imaging. Seminars in Nuclear Medicine, 41(6): 449-462.

Chen H, Zhang J, Xu Y, Chen B, Zhang K. 2012. Performance comparison of artificial neural network and logistic regression model for differentiating lung nodules on CT scans. Expert Systems with Applications, 39(13): 11503-11509.
Kuruvilla J, Gunavathi K. 2014. Lung cancer classification using neural networks for CT images. Computer Methods and Programs in Biomedicine, 113(1):202-209.
Zhao Z, Chen J, Yin X, Song H, Wang X, Wang J. 2015. Establishing assistant diagnosis models of solitary pulmonary nodules based on intelligent algorithms. Cell Physiol Biochem, 35(6):2463-2471.
Sabatier R, Finetti P, Cervera N, Lambaudie E, Esterni B, Mamessier E, Bertucci F. 2011. A gene expression signature identifies two prognostic subgroups of basal breast cancer. Breast Cancer Research and Treatment, 126(2):407-420.
Schrauder MG, Strick R, Schulz-Wendtland R, Strissel PL, Kahmann L, Loehberg CR, Faschin PA. 2012. Circulating micro-RNAs as potential blood-based markers for early stage breast cancer detection. PLoS One, 7(1): e29770.
Giri D, Rajendra AU, Martis RJ, Vinitha SS, Lim TC, Ahamed VT, Suri JS. 2013. Automated diagnosis of Coronary Artery Disease affected patients using LDA, PCA, ICA and Discrete Wavelet Transform. Knowledge-Based Systems, 37(0),:274-282.
Kruppa J, Liu Y, Diener HC, Holste T, Weimar C, König IR, Ziegler A. 2014. Probability estimation with machine learning methods for dichotomous and multicategory outcome: Applications. Biometrical Journal, 56(4):564-583.
Sousa JR, Silva AC, Paiva AC, Nunes RA. 2010. Methodology for automatic detection of lung nodules in computerized tomography images, Computer Methods and Programs in Biomedicine. 98:1–14.
Kim JB, Lee Y, Goo LJ, Kim SS, Kang SH. 2009. Development of an automatic classification system for differentiation of obstructive lung disease using HRCT, Journal of Digital Imaging. 22:136–148.
Keshani M, Azimifar Z, Tajeripour F, Boostani R. 2013. Lung nodule segmentation and recognition using SVM classifier and active contour modeling: a complete intelligent system. Comput Biol Med. 43:287–300.
Madero OH, Vergara VO, Cruz Sanchez VG, Ochoa DJ, Nandayapa AJ. 2015. Automated system for lung nodules classification based on wavelet feature descriptor and support vector machine. Biomed Eng Online, 14(1): 9.
Zinovev D, Feigenbaum J, Furst J, Raicu D. 2011. Probabilistic lung nodule classification with belief decision trees. Paper presented at the Engineering in Medicine and Biology Society, EMBC, 2011 Annual International Conference of the IEEE.
Pradhan B. 2013. A comparative study on the predictive ability of the decision tree, support vector machine and neuro-fuzzy models in landslide susceptibility mapping using GIS. Computers & Geosciences. 51:350-365.
Friedrichs F, Igel Ch. 2005. Evolutionary tuning of multiple SVM parameters. Neurocomputing. 64:107–117.
Huang CL, Wang CJ. 2006. A GA-based feature selection and parameters optimization for support vector machines. Expert Syst Appl. 31:231–240.
Huang CL, Dun JF. 2008. A distributed PSO–SVM hybrid system with features selection and parameter optimization. Appl Soft Comput. 8:1381–1391.
Li Y, Wang J. 2012. A mathematical model for predicting malignancy of solitary pulmonary nodules. World J Surg, 36(4): 830-835.
Mulshine JL, Smith RA. 2002. Lung cancer – 2: Screening and early diagnosis of lung cancer. Thorax. 57(12):1071–1078.
Erasmus JJ, Connolly JE, McAdams HP, et al. 2000. Solitary pulmonary nodules: Part I. Morphologic evaluation for differentiation of benign and malignant lesions. Radiographics. 20:43-58.
Glas AS, Lijmer JG, Prins MH, Bonsel GJ, Bossuyt PM. 2003. The diagnostic odds ratio: A single indicator of test performance. Journal of Clinical Epidemiology. 56(11):1129–1135.