Soft Computing (2022) 26:3369–3383
https://doi.org/10.1007/s00500-021-06718-w

OPTIMIZATION

Improved cost-sensitive multikernel learning support vector machine algorithm based on particle swarm optimization in pulmonary nodule recognition

Yang Li[1] · Jiayue Chang[1] · Ying Tian[2]

Accepted: 8 December 2021 / Published online: 18 January 2022

Abstract

In the lung computer-aided detection (Lung CAD) system, the region of interest (ROI) of lung nodules has more false positives, making the imbalance between positive and negative (true positive and false positive) samples more likely to lead to misclassification of true positive nodules, a cost-sensitive multikernel learning support vector machine (CS-MKL-SVM) algorithm is proposed. Different penalty coefficients are assigned to positive and negative samples, so that the model can better learn the features of true positive nodules and improve the classification effect. To further improve the detection rate of pulmonary nodules and overall recognition accuracy, a score function named F-new based on the harmonic mean of accuracy (ACC) and sensitivity (SEN) is proposed as a fitness function for subsequent particle swarm optimization (PSO) parameter optimization, and a feasibility analysis of this function is performed. Compared with the fitness function that considers only accuracy or sensitivity, both the detection rate and the recognition accuracy of pulmonary nodules can be improved by this new algorithm. Compared with the grid search algorithm, using PSO for parameter search can reduce the model training time by nearly 20 times and achieve rapid parameter optimization. The maximum F-new obtained on the test set is 0.9357 for the proposed algorithm. When the maximum value of F-new is achieved, the corresponding recognition ACC is 91%, and SEN is 96.3%. Compared with the radial basis function in the single kernel, the F-new of the algorithm in this paper is 2.16% higher, ACC is 1.00% higher and SEN is equal. Compared with the polynomial kernel function in the single kernel, the F-new of the algorithm is 3.64% higher, ACC is 1.00% higher and SEN is 7.41% higher. The experimental results show that the F-new, ACC and SEN of the proposed algorithm is the best among them, and the results obtained by using multikernel function combined with F-new index are better than the single kernel function. Compared with the MKL-SVM algorithm of grid search, the ACC of the algorithm in this paper is reduced by 1%, and the results are equal to those of the MKL-SVM algorithm based on PSO only. Compared with the above two algorithms, SEN is increased by 3.71% and 7.41%, respectively. Therefore, it can be seen that the cost sensitive method can effectively reduce the missed detection of nodules, and the availability of the new algorithm can be further verified.

Keywords Multiple kernel learning support vector machine (MKL-SVM) · Particle swarm optimization (PSO) · F-score · Cost-sensitive (CS) · Pulmonary nodules recognition

1 Introduction

Lung cancer has the highest mortality rate in the world. The latest statistics show that in cancers, both the incidence of lung cancer and the death rate of lung cancer rank first for Chinese males. Regarding the statistics in Chinese females, the incidence of lung cancer ranks second, and the death rate of lung cancer ranks first (Cao et al. 2021). According to the number of cases, lung cancer, which increases by approximately 781000 cases every year, ranks first in China. Lung cancer is also the most common type of cancer that kills both men and women in China. According to other statistics, lung cancer is also the main cause of cancer mortality in the USA, accounting for 22% of all male cancer deaths and 22% of all female cancer deaths (Siegel et al. 2021). Early surgery is

✉ Ying Tian
326329416@qq.com

[1] School of Computer Science and Engineering, Changchun University of Technology, Jilin 130012, China

[2] Changchun Obstetrics-Gynecology Hospital, Jilin 130042, China

the most effective treatment for lung cancer, but most patients are diagnosed in the advanced stages of the disease. In 2015, the European Imaging Society and the European Respiratory Society published the latest white paper on lung cancer screening in the European Respiratory Journal (ERJ) , aiming to guide clinical lung cancer screening to facilitate the early detection and treatment of lung cancer (Kauczor et al. 2015). Pulmonary nodule detection and benign or malignant recognition, which can improve the survival chances of individuals, play a vital role in the early diagnosis of lung cancer. However, the five-year survival rate of lung cancer patients is only 10%–15% (Stewart et al. 2019). If health management can be strengthened to achieve the early screening, detection and treatment of lung cancer, the cure rate of patients can increase to 65%, which can effectively improve the survival rate of lung cancer patients and prevent them from missing the best treatment opportunity (Stewart et al. 2019; Liu et al. 2017).

Early detection of lung cancer can be achieved effectively by computed tomography (CT) or positron emission tomography (PET). Pulmonary nodules, which are the early manifestations of lung cancer, are round or quasi-round dense shadows with a diameter of less than 30 mm on CT images. The Lung CAD system is a comprehensive application combining medical image processing, computer vision and machine learning technology that quickly and accurately detects and identifies lung nodules from CT or PET images to provide an efficient screening scheme. At the same time, the Lung CAD can also reduce the visual fatigue of radiologists caused by excessive reading, provide auxiliary second opinions for radiologists, and help radiologists improve the detection efficiency and accuracy of benign or malignant pulmonary nodule recognition. Standard Lung CAD systems usually include image preprocessing, pulmonary parenchymal segmentation, segmentation of a candidate nodule ROI or volume of interest (VOI), the calculation and selection of ROI or VOI features, and benign or malignant pulmonary nodule recognition.

In recent years, machine learning (ML) has been widely applied to Lung CAD systems, especially the support vector machine (SVM) algorithm, which has achieved certain success (Orozco et al. 2015; Manikandan and Bharathi 2016; Zhou et al. 2016; Jose et al. 2017; Li et al. 2018; Jia et al. 2020; Tong et al. 2021). Designing suitable kernel functions for a given problem is a common challenge for both SVM and kernel learning (KL) methods. Compared with the accuracy and stability of classification in single kernel learning (SKL), those in multikernel learning (MKL) can be improved. Bucak et al. (2014) comprehensively described the MKL method, and the latest progress in this field is also noted. Wang et al. (2021) investigated a careful exploration of the connection between MKL methods and deep learning methods, and provided valuable references for future research directions. The MKL-SVM algorithm, which is constructed by combining SVM with MKL, can effectively achieve both learning and generalization ability. The MKL-SVM algorithm was used to identify pulmonary nodules through the study conducted by Li et al. (2013, 2018). More specifically, Li et al. (2013) proposed a hybrid kernel SVM algorithm in which the ROI characteristics are first calculated and designed for the recognition of pulmonary nodules so that multiple recognition indicators, such as the accuracy (ACC), sensitivity (SEN), receiver operating characteristic (ROC) curve and area under the ROC curve (AUC), are simultaneously improved. Furthermore, as the swarm intelligence idea was introduced, the MKL-SVM-PSO algorithm was proposed, the fitness value and optimization process under different inertia weights were discussed, and relatively ideal nonlinear dynamic inertia weights were obtained (Li et al. 2018). The experimental results show that the proposed algorithm can converge to the optimal fitness value quickly, effectively and smoothly, which verifies the effectiveness of the algorithm.

In addition, another high-profile technology in machine learning is deep learning; with the rise of deep learning, many researchers in the field of medical image processing have also introduced deep learning to their research (Tajbakhsh and Suzuki 2016; Hongtao et al. 2018; Monkam et al. 2019; Puttagunta and Ravi 2021). Zhao et al. (2018) developed a hybrid convolutional neural networks (CNN) of LeNet and AlexNet, at the same time, the CNN parameter optimization method was explored, and 1018 cases in lung image database consortium and image database resource initiative (LIDC-IDRI) were selected as experimental data. The final accuracy reached 82.2%, which is helpful for the diagnosis of nodules. To improve the accuracy of nodule recognition, Lin et al. (2020) proposed a model using Taguchi method for parameter optimization based on a two-dimensional CNN, and verified the effectiveness of the model in LIDC-IDRI and SPIE-AAPM data sets, respectively. The results showed that the proposed method can improve the accuracy of recognition. In recent years, there has also been interest in combining deep learning methods with traditional machine learning methods. Based on the CNN architecture, Zhang et al. (2019) used multiple deep CNNs to classify lung nodules to build an ensemble learning model and selected 743 medical images in the LIDC-IDRI for experiments: first, deep CNNs with different architectures are used to predict the nodules; then, voting, SVM, decision trees and other methods are selected to achieve the fusion of different prediction results, and finally, the classification of nodules is achieved. Bansal et al. (2020) proposed to extract deep features and handcrafted features using ResNet network and morphological techniques for the lung nodules classification problem, and XGBoost was selected for classification after feature combination, with an experimental accuracy of 88.30%, which is better than the other techniques mentioned. Polat and Danaei Mehr

(2019) proposed to use AlexNet and GoogleNet network as a new 3D-CNN model, combined with SVM as a classifier to achieve lung nodule classification. Experiments showed that combining deep learning methods with SVM can improve model performance, and finally the accuracy can achieve 91.8%.

Deep learning has played an important role in medical image processing in recent years. Because of its excellent learning ability, high fitting ability, and high accuracy, it is an important tool in the field of medical research. However, the training stage of a deep learning model depends on a large amount of medical image data, and it is still not easy to obtain a high-quality data set with accurate annotations (Francisco and Azuaje 2019). At the same time, an increase in training data will also lead to longer training times and more complex processing. The establishment of deep learning models has a high demand for equipment. At present, in most cases, experiments can only be conducted through transfer learning, there is a lack of a deep theoretical foundation in mathematics, and the interpretability is low. Nevertheless, deep learning can still achieve excellent experimental results and is widely used. The traditional SVM algorithm is a typical small sample learning method with a solid statistical theoretical foundation and strong interpretability (Abe 2010). The complexity of the calculation depends on the number of support vectors rather than the dimensionality of the sample space, which avoids the curse of dimensionality and, at the same time, has a better generalization ability. In addition, SVM has low requirements for the experimental environment and is easy to implement. However, SVM is not ideal for multiclassification problems, as it is not easy to achieve large-scale training samples, and SVM is also sensitive to the selection of the parameters in the kernel functions. However, at present, in the case of an insufficient data set and the lack of sophisticated experimental equipment, SVM is still a feasible method.

Therefore, for the problem of lung nodule recognition, this paper selects SVM for nodule classification and improves the kernel function of the SVM and parameter optimization method to achieve better experimental results.

The MKL-SVM algorithm has achieved certain results in Lung CAD recognition, but there are still two problems that need to be solved:

1. In the process of medical image recognition, the positive and negative class samples are often unbalanced. If the same penalty coefficient is used for samples of different classes, the classification hyperplane of SVM will be skewed toward the minority class, which will cause overfitting for samples of the majority class, making the samples of the minority class more likely to be misclassified than the samples of the majority class (Tao et al. 2019). In the Lung CAD system, in order to prevent the missed detection of nodules, a large number of false positives are usually retained in pulmonary nodule ROI segmentation step before recognition, so that the number of false positives in the extracted candidate nodule ROI is much higher than true positives. However, due to the dominant influence of more false positive nodules, the recognition model tends to pay less attention to the true positive nodules during the learning process or even ignore it, which easily leads to the missed detection of true positive nodes. Therefore, it is significant to solve the class imbalance problem to improve the performance of the system.
2. The evaluation indicators are diverse due to different evaluation perspectives in pulmonary nodule recognition, and there is no unified standard. Therefore, it is difficult to take into account multiple indicators to evaluate pulmonary nodule recognition. The accuracy was taken as the evaluation standard, that is, the overall recognition effect was taken as the evaluation standard of the Lung CAD system (Sui et al. 2015). However, the nodule detection rate was ignored. The sensitivity was taken as the evaluation standard of the Lung CAD system, and the nodule detection degree was represented by the sensitivity (Wang et al. 2015; Javaid et al. 2016). However, too many false nodules are easily identified as nodules, and too many misjudgments occur when the pulmonary nodule ROIs are not balanced, causing psychological pressure and additional examination burdens for patients. The ROC curve and AUC were taken as the evaluation indexes by some researchers (Aoyama et al. 2003; Gao et al. 2015; Gonçalves et al. 2017; da Nóbrega et al. 2018). Other researchers have also explored this issue(Orozco et al. 2015; Demir and Yılmaz Çamurcu 2015; de Carvalho Filho et al. 2017; Li et al. 2019; Ye et al. 2020), and ACC, SEN and SPE are used to evaluate the experimental results on the test set. However, there was no evaluation of the objective function of the training model, and only the test results were discussed. (Sui et al. 2015) discussed this problem, giving attention to multiple indicators; for the evaluation of SVM, an $F_{measure}$ function was adopted to evaluate the accuracy and sensitivity of the positive sample classification results, where the closer the function is to 1, the better the classification result. Although there are many classification algorithms that consider multiple evaluation indicators, there is still much potential for research to optimize and improve the MKL-SVM algorithm.

The main contributions of this article are as follows:

1. In the pulmonary nodule recognition problem, a cost-sensitive multikernel learning support vector machine (CS-MKL-SVM) algorithm was proposed for the imbal-

ance of positive and negative samples, which is likely to cause too many missed detections. Different penalty coefficients are used for true positive nodules and false positive nodules to increase the penalty for misclassified samples, so that the model can better learn the features of true positive to improve the recognition results.

2. Since the CS-MKL-SVM algorithm has many parameters, the CS-MKL-SVM algorithm based on PSO is proposed in this paper. The grid search algorithm can be used to find the optimal parameter set in the cross-validation. However, due to the numerous parameters and oversized nested loops and levels, the number of calculations and training time are indeed too long in the CS-MKL-SVM algorithm. In addition, if the grid search step size is not dense enough, the global optimal solution is difficult to find. However, a swarm intelligence algorithm can find the optimal parameter set quickly without traversing all the parameter points. Hence, in this paper, the CS-MKL-SVM algorithm is proposed based on PSO.
3. To better balance the overall accuracy and the detection rate of nodules, to prevent the missed detection of nodules, this paper proposes a F-new function as the PSO evaluation standard for parameter optimization in the CS-MKL-SVM algorithm. Moreover, in this algorithm, F-new under the cross-validation is used as the fitness function value to replace the traditional ACC or SEN single-indicator evaluation system. The theoretical analysis of the F-new function proves that the function can take into account both ACC and SEN, and the experimental results further show the feasibility of this index. The algorithm in this paper can ensure the accuracy of the overall nodule recognition and at the same time improve the detection rate of nodules.

The rest of the article is organized as follows. Section 2 introduces some related work on the CS-SVM algorithm. In Section 3, the proposed framework is specified. Section 4 shows the experimental results of the experiments in this paper. Finally, a conclusion of the work is shown in Section 5.

2 CS-MKL-SVM Algorithm for Unbalanced Data Sets

2.1 Cost-Sensitive MKL-SVM Algorithm

The same penalty parameters, which are the regularization coefficients named C, are used in the standard SVM for positive and negative samples. The values of the penalty parameters determine how the maximum class interval and the minimum training error compromise in SVM. However, the classification will be skewed toward the side containing fewer samples when the positive and negative sample data sets are seriously unbalanced in SVM, thus affecting the classification accuracy. For example, if the same penalty coefficient C is used when the positive sample is small, the sum of errors corresponding to the positive samples (cost function or loss function) will be less than the sum of errors corresponding to the negative samples (cost function or loss function). In other words, a large penalty coefficient is applied to the negative samples so that the segmentation sample plane moves to the side of the positive samples. Based on the MKL-SVM algorithm, a cost-sensitive MKL-SVM (CS-MKL-SVM) algorithm is proposed to solve the positive and negative sample imbalance. Different penalty coefficients C_+ and C_- are introduced to the positive and negative samples to flexibly adjust the misclassification costs of the false positives and false negatives.

The training pattern is expressed as:

$$T = \{(x_i, y_i)\} \in (X \times Y)^l \tag{1}$$

where the parameter l is the number of training patterns. The parameter x_i is the input vector of SVM, and the parameter $y_i \in \{-1, +1\}$ is the category label. In this paper, $x_i \in R^n$ corresponds to the eigenvector that is extracted from the i^{th} sample from the n-dimensional ROI. The value of parameter $y_i = 1$ corresponds to the positive samples, which are the true positive nodules and are malignant tumor lesions. The value of parameter $y_i = -1$ denotes negative samples, which are false nodules or false tumors.

The mathematical model of the original CS-MKL-SVM algorithm is described as follows:

$$\begin{cases} \min \ \frac{1}{2}\|w\|^2 + C_+ \sum\limits_{y_i=+1} \xi_i + C_- \sum\limits_{y_i=-1} \xi_i \\ s.t. \ \sum\limits_{i=1}^{l} y_i \left(w \cdot \Phi\left(x_i\right) + b\right) \geq 1 - \xi_i \\ \qquad \xi_i \geq 0 \\ \qquad i = 1, 2, \ldots, l \end{cases} \tag{2}$$

The parameter ξ_i is the slack variable, b is the constant bias, and $\Phi(x_i)$ is the input sample mapping in the high-dimensional feature space. The original CS-MKL-SVM problem is transformed into a Lagrangian dual problem by a Lagrange multiplier:

$$\begin{cases} \max \ \sum\limits_{i=1}^{l} \alpha_i - \frac{1}{2} \sum\limits_{i=1}^{l} \sum\limits_{j=1}^{l} \alpha_i \alpha_j y_i y_j K\left(x_i, x_j\right) \\ s.t. \ \sum\limits_{i=1}^{l} \alpha_i y_i = 0 \\ \qquad 0 \leq \alpha_i \leq C_+, y_i = +1 \\ \qquad 0 \leq \alpha_i \leq C_-, y_i = -1 \end{cases} \tag{3}$$

where

$$K(x_i, x_j) = \Phi(x_i) \cdot \Phi(x_j) \tag{4}$$

The discriminant function is as follows:

$$f(x) = \operatorname{sgn}\left(\sum_{i=1}^{l} \alpha_i y_i K(x_i, x) + b\right) \tag{5}$$

where the function $K(x_i, x_j)$ is the kernel function involving only the inner product of the input sample vectors, and the value is constant. The kernel function is a symmetric function, and two n-dimensional real eigenvectors are converted into a real number using the kernel function, i.e., $K : (R^n \times R^n \rightarrow R)$. As long as the Mercer condition is satisfied, the kernel function $K(x_i, x_j)$ will take the dot product of a feature space. One of the keys to improving the performance of SVM is to design suitable kernel functions for a given problem. The common kernel functions include the polynomial kernel function and radial basis function (RBF) which are expressed as K_{poly} and K_{rbf} :

$$K_{poly}(x, y) = (x^t y + 1)^d \tag{6}$$

$$K_{rbf}(x, y) = exp(-\|x - y\|^2/2g^2) \tag{7}$$

where the parameter d represents the polynomial kernel degree and g represents the RBF kernel width. The parameters d and g are superparameters that need to be given in advance. The RBF kernel has a strong learning ability, and the polynomial kernel has a strong generalization ability; thus, the combination of the two can take into account the abilities of both learning and generalization. It has been proven that equation (8) is still a kernel function and that the combination of the RBF kernel and polynomial kernel is a convex combination (Li et al. 2018).

$$\begin{cases} K_{mix}\left(x_i, x_j\right) = mK_{poly}\left(x_i, x_j\right) + (1-m)\, K_{rbf}\left(x_i, x_j\right) \\ 0 < m < 1 \end{cases} \tag{8}$$

where m is the weight coefficient of the polynomial kernel function, and the proportion of each basic kernel function in the multikernel function can be adjusted freely by m. The parameter K_{mix} in equation (8) is taken as the multikernel function K in the CS-MKL-SVM algorithm.

The CS-MKL-SVM parameters are numerous, including the regularization coefficients C_+ and C_- of the positive and negative samples, the polynomial kernel degree d, the RBF kernel width g, and the multikernel weight m. The grid search algorithm can be used to find the optimal parameter set in the sense of cross-validation (CV). However, due to the excessive numbers of parameters in the nested cycles and layers, grid searching leads to a large number of calculations and a long running time. Moreover, the global optimal solution cannot be found if the grid search step size is not dense enough. However, the optimal parameter set can be found quickly by a heuristic algorithm without traversing all parameter points.

2.2 CS-MKL-SVM Optimization Algorithm based on PSO

Particle swarm optimization (PSO), which searches by tracking the optimal particle in the solution space, is a heuristic algorithm based on swarm intelligence.

Suppose that in a D-dimensional search space, the species group $X = (X_1, X_2, \ldots, X_n)$ is composed of n particles, and $X_i = (x_{i1}, x_{i2}, \ldots, x_{iD})^T$ represents the position of the i^{th} particle in the D-dimensional parameter searching space. The fitness value corresponding to each particle X_i can be calculated according to the objective function. The velocity of the ith particle is $V_i = (V_{i1}, V_{i2}, \ldots, V_{iD})^T$, its individual extremum is $P_i = (P_{i1}, P_{i2}, \ldots, P_{iD})^T$, and the group extremum is $P_g = (P_{g1}, P_{g2}, \ldots, P_{gD})^T$. In each iteration, the particle updates its speed and position through the individual extremum and group extremum, and the updated expression is as follows:

$$V_{id}^{k+1} = \omega V_{id}^{k} + c_1 r_1 (P_{id}^{k} - X_{id}^{k}) + c_2 r_2 (P_{gd}^{k} - X_{id}^{k}) \tag{9}$$

$$X_{id}^{k+1} = X_{id}^{k} + V_{id}^{k+1} \tag{10}$$

where k is the current number of iterations, ω is the inertia weight, $d = 1, 2, \ldots, D$, V_{id} is the velocity of the i^{th} particle in the D-dimensional space,c_1 and c_2 are nonnegative constant acceleration factors, and the parameters r_1 and r_2 are random numbers distributed in the interval [0, 1]. To prevent the blind search of particles, the velocity and position are usually limited to $[-V_{max}, V_{max}]$ and $[-X_{max}, X_{max}]$ respectively.

2.3 The New Target Function based on the *F*-score

In the classic PSO algorithm, a single evaluation metric in the sense of CV is usually taken as the final goal and is determined as the fitness function value, but it cannot evaluate the recognition results well for special problems. Therefore, the concept of harmonic mean in statistics is introduced. In statistics, the harmonic mean is the inverse of the arithmetic mean of the inverse of each statistical variable (Parkash and Thukral 2010), as shown in equation (11):

$$H = \frac{n}{\frac{1}{x_1} + \frac{1}{x_2} + \cdots + \frac{1}{x_n}} \tag{11}$$

where H is the harmonic mean, n is the total number of variables, and x_i represents the i^{th} variable, $i = 1, 2, \ldots, n$.

In some systems of machine learning, $Precision$ and $Recall$ are usually the more concerned metrics. Among them, $Precision$ is how many of all samples that the model determines as positive are really positive samples; $Recall$ is how many of all positive samples are determined as positive by the model, that is, the detection rate of positive samples. The two metrics are shown in equation (12) and equation (13), respectively.

$$Precision = \frac{TP}{TP + FP} \tag{12}$$

$$Recall = \frac{TP}{TP + FN} \tag{13}$$

where true positive (TP) represents the number of positive classes predicted as positive; true negative (TN) represents the number of negative classes predicted as negative; false positive (FP) represents the number of negative classes predicted as positive; and false negative (FN) represents the number of positive classes predicted as negative.

In some specific problems, $Precision$ and $Recall$ are both important and need to be considered comprehensively. As an important metric to weigh $Precision$ and $Recall$, the $F_{measure}$ function represents the weighted harmonic mean of $Precision$ and $Recall$ (Sui et al. 2015), as shown in equation (14):

$$\begin{aligned} F_{measure} &= \frac{1}{\alpha \frac{1}{Precision} + (1-\alpha)\frac{1}{Recall}} \\ &= \frac{(\beta^2 + 1) Precision * Recall}{\beta^2 Precision + \text{Recall}} \end{aligned} \tag{14}$$

where $\beta^2 = \frac{1-\alpha}{\alpha}$, $\alpha \in [0, 1]$, $\beta^2 \in [0, \infty]$. When $\beta > 1$, $Precision$ has a greater impact; when $\beta < 1$, $Recall$ has a greater impact; when $\beta = 1$, it is the $F_1 - score$ function, which means that $Precision$ and $Recall$ have the same importance, as shown in equation (15):

$$F_1\text{−score} = \frac{2 * Precision * Recall}{Precision + Recall} \tag{15}$$

The standard $F_1 - score$ function was earlier proposed mainly for the information retrieval problems (Rijsbergen 1974; Guns et al. 2012). In information retrieval, it is more concerned about " what percentage of the retrieved information is of interest to the users ($Precision$) " and " what percentage of the content of interest to the users is retrieved ($Recall$)"; therefore, $Precision$ and $Recall$ are more suitable as performance metrics for such requirements.

However, in the Lung CAD system, the ACC represents the overall correct recognition probability. The $Recall$ indicator is also called the sensitivity (SEN) indicator in the Lung CAD system, which represents the detection probability of true positive nodules. The higher SEN value, the fewer false negative nodules, and the less possibility of missed detection. $Precision$ represents the ability to recognize non-nodules. The higher the $Precision$ value, the fewer false positive nodules, and the less probability of incorrectly diagnosed as nodules. For Lung CAD, on the one hand, the overall recognition effect is very important. On the other hand, in order to prevent the occurrence of the missed detection, the recognition rate of malignant lesions cannot be ignored, that is, the SEN indicator, which is the detection rate of malignant tumors. The definition equations of ACC and SEN are shown in equation (16) and equation (17), respectively.

$$ACC = \frac{TP + TN}{TP + TN + FP + FN} \tag{16}$$

$$SEN(Recall) = \frac{TP}{TP + FN} \tag{17}$$

In the above formula, TP is the number of true positive nodules detected, FP is the number of false positive nodules detected, FN is the number of false negative nodules detected, and TN is the number of true negative nodules detected or the number of false nodules.

For the recognition of lung nodules, more attention is paid to " the accuracy of overall nodules recognition (ACC) " and " the true detection rate of lung nodules (SEN)." The missed detection of nodules may cause the patient to miss the optimal time for treatment. For a prediction model, the higher SEN is, the smaller the FN and missing nodule detection rates are; thus, the detection result will be more trusted. Therefore, SEN is as important as ACC. In this paper, for the evaluation criteria of the Lung CAD, the mentioned F-new function is shown in equation (18), F-new is used as the fitness function of the CS-MKL-SVM-PSO algorithm in the sense of cross-validation, and then the evaluation criteria for parameter optimization are established as a substitute the traditional single ACC evaluation system, taking into account the overall recognition accuracy and reducing the missed detection of nodules.

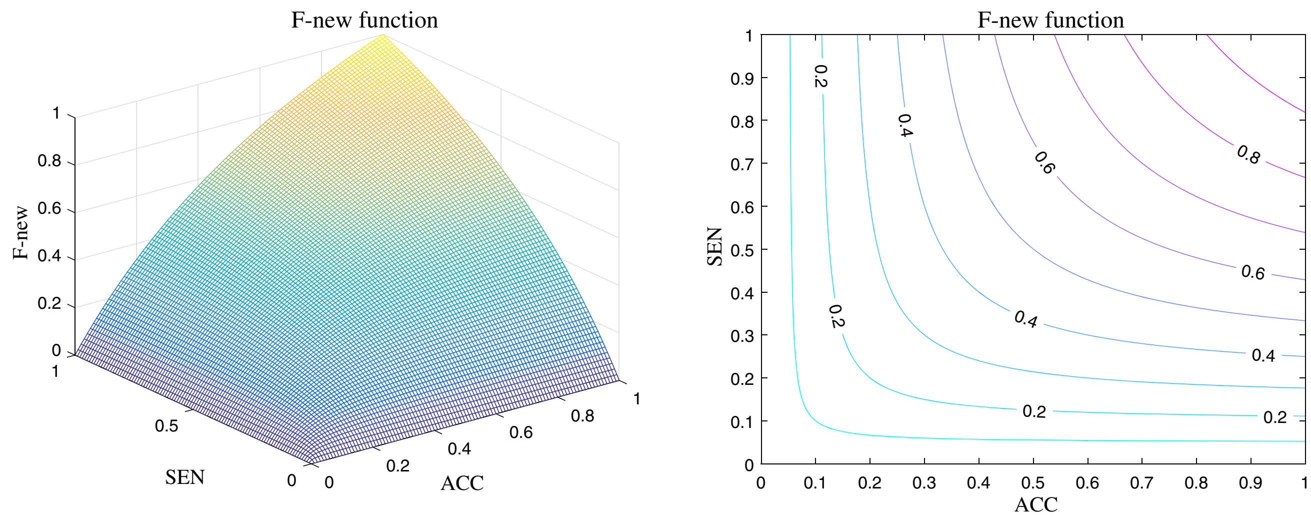
$$F - \text{new} = \frac{2}{\frac{1}{ACC} + \frac{1}{SEN}} = \frac{2 * ACC * SEN}{ACC + SEN} \tag{18}$$

The F-new function is the harmonic mean of ACC and SEN.

Conclusion 1 The value range of ACC and SEN are both [0,1]. When ACC and SEN are equal to 1, the F-new function can obtain the maximum value, and the maximum value is 1.

Proof The reference have proved that formula (19) is established (Liao and Wu 2015).

$$\frac{x_1 + x_2}{2} \geq \sqrt{x_1 x_2} \geq \frac{2}{\frac{1}{x_1} + \frac{1}{x_2}} \tag{19}$$



(a) The 3-dimensional view

(b) The contour lines

Fig. 1 F-new function simulation diagram

In the above formula, the equal sign holds if and only if $x_1 = x_2$. It is well known that $\frac{x_1+x_2}{2}$ is an increasing function. When $x_1 \in [0, 1], x_2 \in [0, 1]$, $\frac{x_1+x_2}{2}$ takes the maximum value is 1 at $x_1 = x_2$. Therefore, $\frac{2}{\frac{1}{x_1}+\frac{1}{x_2}}$ takes the maximum value is 1 at $x_1 = x_2$.

Using ACC and SEN as input, the simulation experiment is conducted on the harmonic average F-new, and the simulation results of the three-dimensional view and contour lines obtained are shown in Fig. 1. □

As shown in the three-dimensional view of Fig. 1(a), ACC and SEN are used as inputs, and the F-new function increases monotonically with the increase in ACC and SEN indicators. The relevant theory can also be proved as follows:

As shown above

$$F_{new}(ACC, SEN) = \frac{2}{\frac{1}{ACC} + \frac{1}{SEN}} = \frac{2ACC \cdot SEN}{ACC + SEN} \quad (20)$$

where $ACC \in [0, 1]$, $SEN \in [0, 1]$, calculate the partial derivatives of ACC and SEN, respectively, as shown in equation (21) and equation (22) :

$$\frac{\partial F_{new}(ACC, SEN)}{\partial ACC} = \frac{2SEN^2}{(ACC + SEN)^2} \geq 0 \quad (21)$$

$$\frac{\partial F_{new}(ACC, SEN)}{\partial SEN} = \frac{2ACC^2}{(ACC + SEN)^2} \geq 0 \quad (22)$$

From the derivative, the first-order partial derivative of $F_{new}(ACC, SEN)$ with respect to ACC is greater than or equal to 0, and $F_{new}(ACC, SEN)$ increases with the increase in ACC. Similarly, $F_{new}(ACC, SEN)$ increases with the increase in SEN. Therefore, when $ACC \in [0, 1]$, $SEN \in [0, 1]$, $F_{new}(ACC, SEN)$ is a monotonically increasing function.

When both ACC and SEN are close to 1, the F-score reaches its maximum, proving that both ACC and SEN are well-considered. As seen from the contour lines of Fig. 1(b), along the direction from (0,0) to (1,1), the closer to the upper right corner, the higher the score of the F-new function. The F-new function gives higher scores to the models in which ACC and SEN are closer, and combined with the previous conclusion, F-new increases monotonically with SEN and ACC indicators, ACC and SEN are increasing and bounded, so when the ACC and SEN are larger and closer, the F-new score is higher. The extreme value is when $ACC = SEN = 1$, F-new = 1 can be obtained , that is, the upper right vertex. The larger the interval between the ACC and SEN is, the lower the F-new score will be. Hence, the model almost loses its practical value.

3 CS-MKL-SVM-PSO Algorithm for Pulmonary Nodule Recognition in Lung CAD

3.1 The Process of the CS-MKL-SVM-PSO Algorithm for Pulmonary Nodule Recognition in Lung CAD

The overall process of the Lung CAD system is shown in Fig. 2. The Lung CAD system is divided into three parts. The first part is the preparation for benign and malignant pulmonary nodule recognition, including the preprocessing of CT images, segmentation of pulmonary parenchyma, seg-

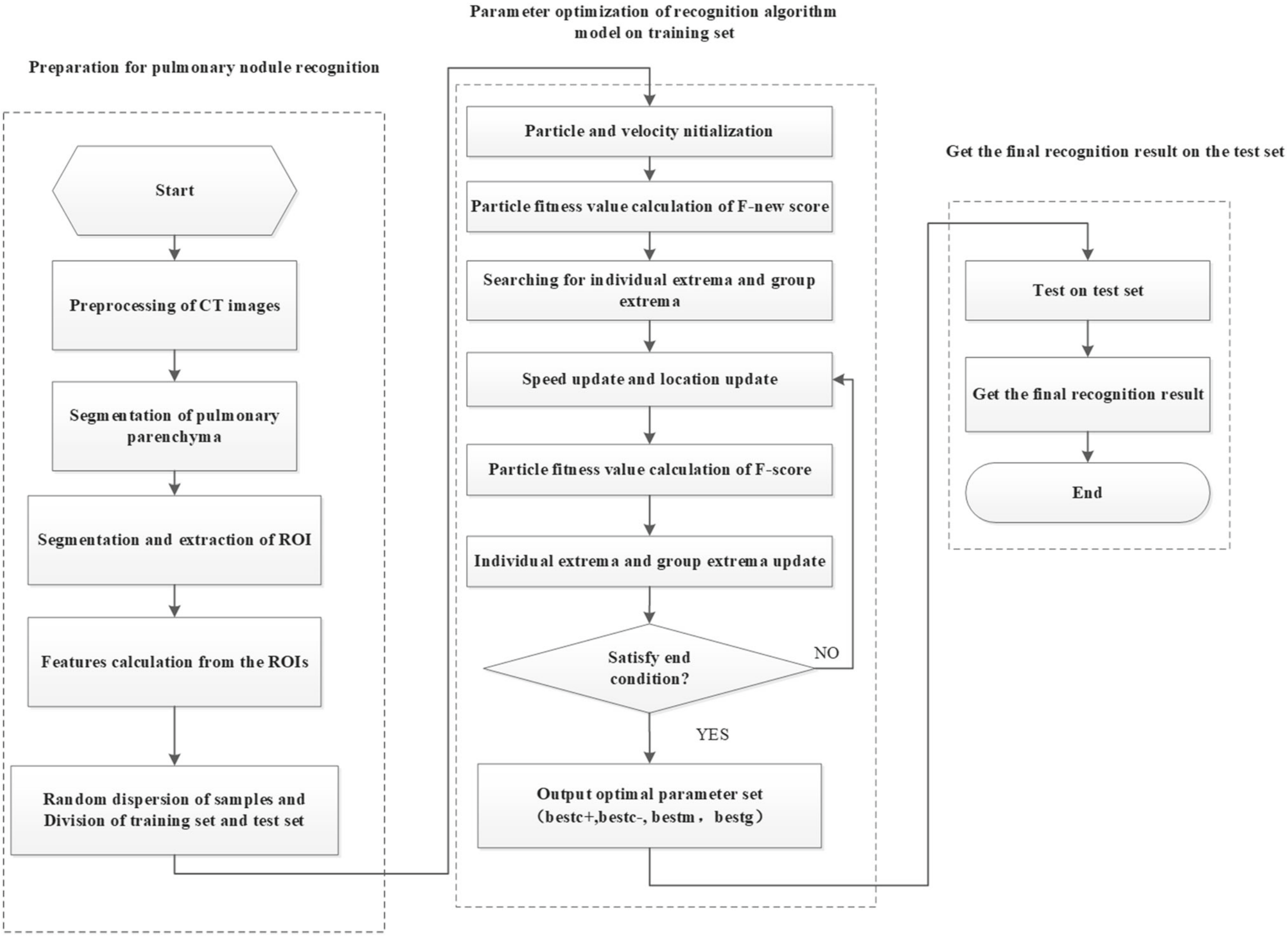

Preparation for pulmonary nodule recognition
Start
Preprocessing of CT images
Segmentation of pulmonary parenchyma
Segmentation and extraction of ROI
Features calculation from the ROIs
Random dispersion of samples and Division of training set and test set
Parameter optimization of recognition algorithm model on training set
Particle and velocity nitialization
Particle fitness value calculation of F-new score
Searching for individual extrema and group extrema
Speed update and location update
Particle fitness value calculation of F-score
Individual extrema and group extrema update
Satisfy end condition?
NO
YES
Output optimal parameter set (bestc+,bestc-, bestm， bestg)
Get the final recognition result on the test set
Test on test set
Get the final recognition result
End

Fig. 2 Flowchart of CS-MKL-SVM-PSO algorithm for pulmonary nodules recognition in Lung CAD

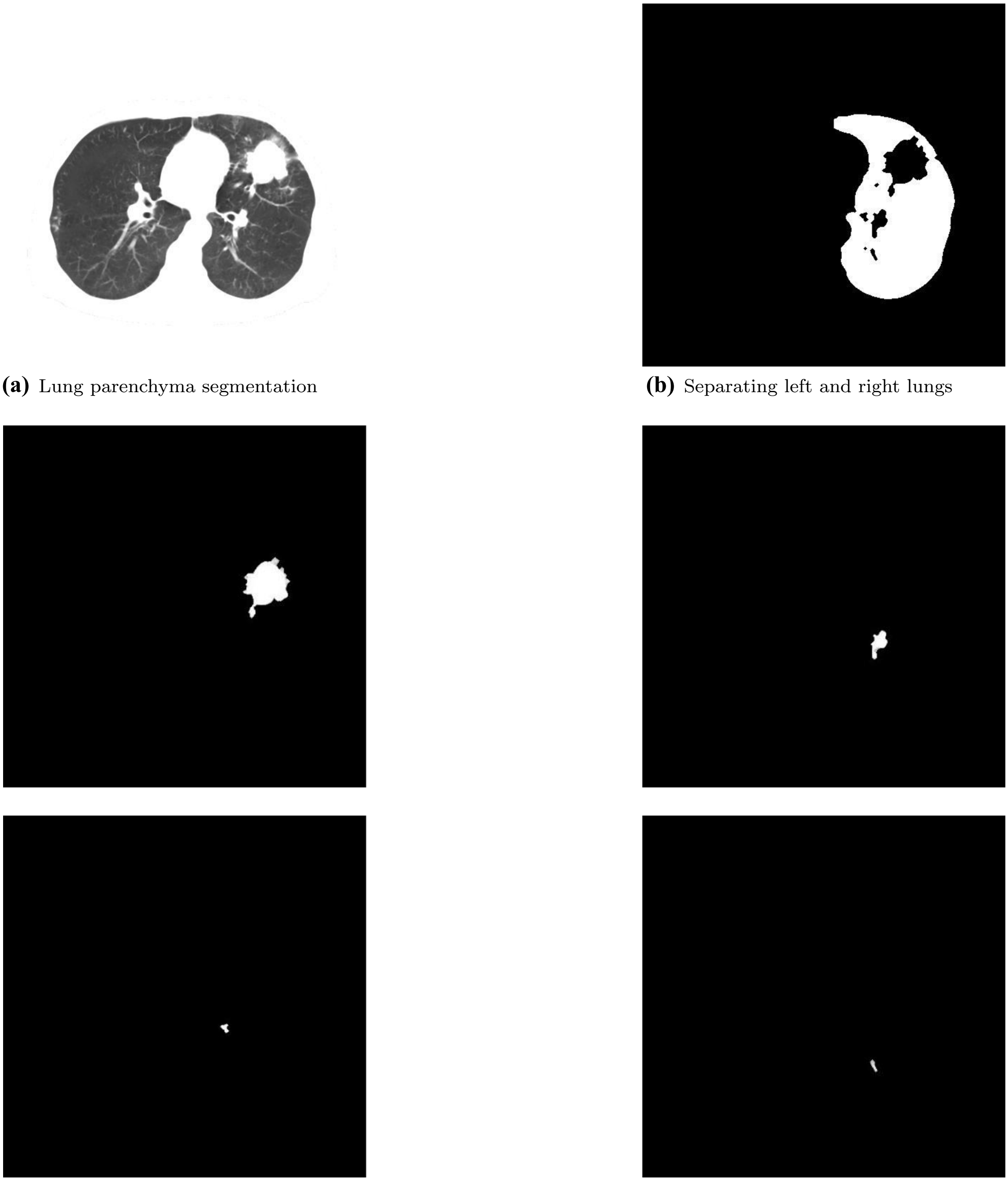
mentation and extraction of ROIs, and feature calculation from the ROIs. The second part is to obtain the recognition algorithm model by quickly optimizing the parameters on the training set with the CS-MKL-SVM-PSO algorithm. The last part is to obtain the final recognition result on the test set. Each of the three parts is indicated by a dotted box.

3.2 Preparation for Pulmonary Nodule Recognition

The shape and structure of pulmonary nodules and their gray features in CT images are different, which makes them easy to misdiagnose and falsely detect. Pulmonary nodule recognition as well as the core module of the Lung CAD is the final step. The selected recognition algorithm will directly affect the detection results.

During image preprocessing, binarization processing on the original CT image, and the largest eight connected regions are reconstructed to remove the background and obtain the pulmonary parenchyma region. Because nodules grow in the pulmonary parenchyma area, segmentation of the pulmonary parenchyma is mainly used to separate the left and right lungs and then extract the lung edges. After the segmentation of the lung parenchyma is completed, further segmentation is needed to extract the relevant ROIs. In this paper, the ROI was segmented by image enhancement technology. To highlight the ROI of the candidate pulmonary nodules, the gray level of the image obtained in the ROI was the same as that of the original image after image reconstruction. Figure 3 illustrates the complete process of segmenting and extracting pulmonary nodule ROIs from a complete original CT image through pulmonary parenchyma segmentation. Due to space limitations, only the pulmonary parenchyma on one side containing pulmonary nodules is given. In the actual experiment, all segments of the pulmonary parenchyma are traversed.

In each case, several pulmonary nodule ROIs can be divided, but there is only one corresponding true positive, namely pulmonary nodules (malignant lesions), and the rest are false positive. This is because CT images are tomography images, which often contain points where blood vessels intersect and end, and their shapes are similar to those of nodules, which are also round. It can be seen from the extracted



(a) Lung parenchyma segmentation

(b) Separating left and right lungs

(c) The 4 pulmonary nodules ROI extracted from the cases

Fig. 3 A complete lung nodule ROI extraction process from original CT image

ROI that the number of false positives is larger than the actual number of lesions, resulting in a serious imbalance in the data set, which will affect the subsequent recognition results. To address this issue, in this paper, a cost-sensitive multikernel support vector machine (CS-MKL-SVM) algorithm is proposed.

The feature selection algorithm has been mentioned in a previous article (Li et al. 2013). Seven morphological features, two grayscale features and four texture features of the ROI are selected, and a total of 13 feature parameters need to be calculated. The seven morphological features selected are area, diameter, circumference, rectangularity, flattening, roundness, and slenderness. Gray features are measured by the gray mean and gray variance. The four texture features are energy, contrast, entropy and inverse difference moment.

Table 1 Twenty times experimental statistics in the training stage

	Poly order	Fitness (objective function) value					Training time (s)			
Experiment	d	Fitness function	Max	Min	Mean	Median	Max	Min	Mean	Median
CS-MKL-SVM-PSO	2	F-new	0.9048	0.8874	0.8960	0.8927	651	461	504	491
	3	F-new	0.9076	0.8744	0.8993	0.9048	557	490	514	512
algorithm A	2	ACC	0.9412	0.9353	0.9379	0.9353	401	364	371	366
algorithm B	2	ACC	0.9412	0.9412	0.9412	0.9412	21493	8028	11270	8210

Table 2 Experimental results of algorithm test stage in this paper

	d	F_{max}	F_{min}	F_{mean}	F_{median}	$MFACC$	$MFSEN$
CS-MKL-SVM-PSO	2	0.9304	0.8753	0.8915	0.8753	90.00%	96.30%
	3	0.9357	0.8753	0.9112	0.9196	91.00%	96.30%

3.3 Model Parameter Optimization

After the above preparation for pulmonary nodule recognition is completed, the training phase of the model is begun. Due to the overfull parameter quantities and excessive optimization time of the CS-MKL-SVM algorithm, we introduce PSO to realize rapid parameter optimization, namely CS-MKL-SVM-PSO. The F-new function under fivefold cross-verification is determined as the fitness function of the CS-MKL-SVM-PSO algorithm. First, the particle species group and velocity are initialized provide the maximum number of iterations maxgen is 200, the species group number is 20, and the number of cross-verification folds is 5. The expression $X_i = (x_{i1}, x_{i2}, x_{i3}, x_{i4})^T$ of each particle respects the solution of the ith $(i = 1, 2, \ldots, 20)$ particle, meaning D=4. The parameters x_{i1}, x_{i2}, x_{i3} and x_{i4} need to be searched, corresponding to the RBF kernel width g, multikernel weight m and regularization coefficients C_+ and C_- of positive and negative class samples, respectively. In equations (9)and(10), ω is the inertia weight, k is the current number of iterations, V_{id} is the velocity of the particles, and the acceleration factors are provided as c_1=1.5 and c_2=1.7. To prevent blindly searching for a particle, its position and velocity are usually restricted to $[-X_{\max}, X_{\max}]$ and $[-V_{\max}, V_{\max}]$, respectively. The position and velocity of each particle are set as follows: the value range of parameter g is $[2^{-7}, 2^7]$, and the velocity is $[-2^9 * 0.6, 2^9 * 0.6]$; the value range of m is [0, 1], and the velocity is [−0.6, 0.6]; the value range of parameter C_+ is $[2^{-9}, 2^9]$, and the velocity is $[-2^9 * 0.6, 2^9 * 0.6]$; the value range of parameter C_- is $\left[2^{-9}, 2^{11}\right]$, and the speed is $\left[-2^{11} * 0.6, 2^{11} * 0.6\right]$.Then, the F-new value is calculated, and the individual extreme value and group extreme value according to the F-new index of the new population particles are updated. When the maximum number of iterations is reached, the update is stopped, and the optimal parameter set is output. Finally, the trained model is tested on the test set to obtain the final recognition result.

4 Pulmonary Nodules Recognition Experiment in Lung CAD

4.1 Experimental Data and Platform

Experimental data were obtained from large specialized hospitals in Jilin Province, China. The experiment selected 20 cases from the case database, with a total of approximately 700 CT tomography images (sections) and an average of approximately 35 images for each case. Each case was equipped with the diagnostic criteria and pathological analysis report from a doctor. Each CT slice image was 512 × 512 pixels in size and 5.0 mm in thickness. After the previous preparation steps for pulmonary nodule recognition from approximately 700 CT images from 20 patients, a total of 270 pulmonary nodule ROIs were extracted, including 80 nodules (malignant lesions) and 190 false positives. A true positive nodule here refers to a malignant tumor according to the pathological analysis report and diagnosis. After the feature selection, the data samples were randomly divided into two groups: 170 training samples (with 53 nodules and 117 false nodules) and 100 test samples (with 27 nodules and 73 false nodules).

Using MATLAB as the experimental platform, the above calculated 13-dimensional characteristic data were normalized by the mapminmax function. The simulation experiment was conducted using the LIBSVM toolbox which can be obtained freely (Chang and Lin 2011). In the model training stage, a fivefold cross-validation was used to evaluate the F-new function which was selected as the parameter optimization standard of the CS-MKL-SVM-PSO algorithm as shown in equation (18).

4.2 Experimental Results Analysis

4.2.1 Experimental results analysis during training stage

In Table 1, the results obtained during the training stage by the multikernel functions with degrees d=2 and d=3 are listed. Moreover, the results of the proposed algorithm are also compared with the results of two algorithms: algorithm A (Li et al. 2018) is the MKL-SVM-PSO algorithm with inertial weight = 1 , and algorithm B (Li et al. 2013) is the MKL-SVM algorithm based on grid search. To ensure the experimental reliability, each experiment was repeated 20 times. The fitness values of the proposed algorithm, algorithm A and algorithm B and their statistical values are listed on the left side in Table 1, and the statistical results of the training time are listed on the right side.

Note that in the algorithm A and algorithm B, the fitness function or objective function searches for the maximum ACC in the training set, which is the overall recognition of ACC, while the proposed F-new function, which comprises the harmonic mean of ACC and SEN, is adopted as the corresponding fitness function. Hence, the fitness function value of the proposed algorithm is less than that of algorithm A and algorithm B. In this paper, the CS-MKL-SVM-PSO algorithm is adopted, and the fitness function is F-new. Under the same value of ACC, according to equation (18) , when SEN is less than or equal to 1, F-new is less than or equal to ACC; if and only if SEN=1, when equivalence is achieved it means there is no missing detection.

The above training results show that the algorithm finds the optimal parameter group when the polynomial kernel has an order d=3. The maximum value of F-new in the training stage is 0.9076, the minimum value is 0.8744, the mean value is 0.8993, and the median value is 0.9048, all of which are higher than those in the case of d=2. The training time of the proposed algorithm is on the same order of magnitude as that of algorithm A, whereas algorithm B obtains the optimal parameter group of the grid search algorithm through the use of finer grid division, and hence, the training time is almost 20 times that of the algorithm in this paper.

4.2.2 Experimental results analysis of the multiple kernel function and single kernel functions

Table 2 shows the experimental results obtained by the algorithm in this paper on the test set along with the statistical results obtained from the F-new function in the test stage. F_{max} and F_{min} represent the maximum and minimum values of the F-new function in 20 experiments, F_{mean} and F_{median} represent the statistical mean and median of the F-new function, and $MFACC$ and $MFSEN$ represent the ACC and SEN values corresponding to the maximum F-new function, respectively. The experimental results showed that the test

Table 3 Experimental results of different kernel functions in the testing stage

Experiment	Parameter	Testing result									
	d	F_{mean}	F_{max}	MA_{mean}	MA_{median}	MA	$MASEN$	MS_{mean}	MS_{median}	MS	$MSACC$
CS-MKL-SVM-PSO	3	0.9112	0.9357	88.95%	89.00%	92.00%	88.89%	93.52%	96.30%	96.30%	91.00%
Polynomial kernel	2	0.8666	0.8944	90.40%	90.00%	91.00%	85.19%	87.41%	88.89%	88.89%	90.00%
Polynomial kernel	3	0.8993	0.8993	91.00%	91.00%	91.00%	88.89%	88.89%	88.89%	88.89%	91.00%
RBF kernel	-	0.9045	0.9141	87.80%	87.00%	91.00%	88.89%	93.33%	92.59%	96.30%	87.00%

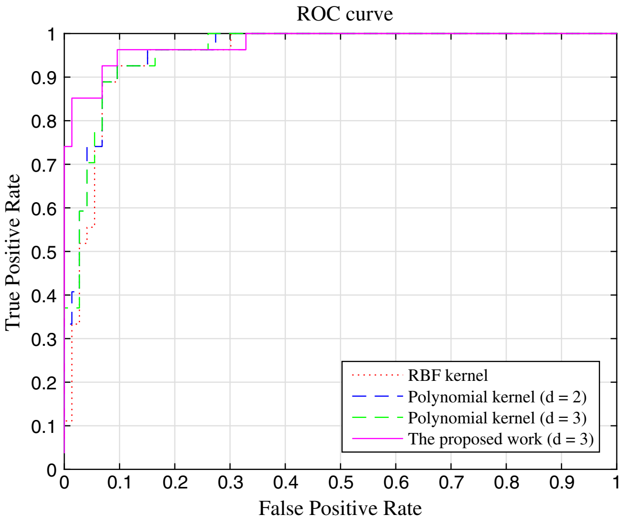


Fig. 4 ROC curves of different kernel functions

results are optimal when d=3; at this point, F_{max} is 0.9357, F_{mean} is 0.9112, F_{median} is 0.9196, $MFACC$=91%, and $MFSEN$=96.3%. Regardless of the observation from the training stage and testing stage, the learning ability and generalization ability are better at the multikernel function d=3, which is the optimal algorithm in this paper.

Further, the PSO algorithm is used as the optimization algorithm and the proposed F-new function is used as the fitness function. The polynomial function and RBF kernel in the single kernel function and the multiple kernel function algorithm in this paper are used for experiments. Table 3 lists the statistical mean values of the test results of different kernel functions during the testing stage, and the ROC curves are shown in Fig. 4. Table 4 showed the AUC for each of the different kernel functions, and the larger the AUC value, the better the classifier effect.

In Table 3, MA represents the maximum ACC obtained by 20 experiments on the test set, and $MASEN$ represents the SEN test index corresponding to MA. In the 20 experiments, the mean and median values of ACC are expressed by MA_{mean} and MA_{median}, respectively. Similarly, MS represents the maximum SEN obtained on the test set, and $MSACC$ represents the ACC indicator corresponding to MS. In 20 times experiments, the mean and median values of SEN are expressed by MS_{mean} and MS_{median}, respectively.

As shown in Table 3, compared with the single kernel functions, the F_{mean} of the proposed multiple kernel function algorithm reaches 0.9112, which is higher than the polynomial kernel function and RBF kernel function. When the polynomial single kernel function is used, the result is better at d=3 with MA_{mean} of 91%, but the SEN value is lower, MS_{mean} of 88.89%. When the RBF kernel function is used, the SEN is higher, MS_{mean} of 93.33%, but the ACC value is lower, MA_{mean} of 87.80%. When using the algorithm in this paper, compared with the polynomial kernel function, although MA_{mean} is reduced by 2.05%, MS_{mean} is increased by 4.63%, which can reduce the missed detection of nodules. Compared with the RBF kernel function, ACC and SEN are both improved. As shown in Fig. 4, the ROC curve represents the ratio of the true positive rate to the false positive rate. The closer the upper left vertex is to the (0,1) point, the higher the AUC value, the better the classifier performance. The upper left vertex of the ROC curve of the algorithm in this paper is closer to the (0,1) point, the AUC is 0.9777, which has better classification performance. Therefore, the MKL method can improve the classification performance of the model compared with the single kernel function.

4.2.3 Experimental results analysis of introducing cost-sensitive and F-new function

Further comparisons are made between the experimental results of algorithm A and algorithm B and the experimental results of the algorithm in this paper on the test set. Table 5 lists the statistical values of the 20 test results obtained by the algorithm in this paper and the related algorithms on the test set.

Algorithm A is the MKL-SVM-PSO algorithm, which uses the swarm intelligence algorithm for parameter optimization and can greatly reduce the training time of the model, but the PSO algorithm tends to lose the particle diversity in the late iteration, so that the global optimal solution cannot be obtained. Algorithm B is the MKL-SVM algorithm, which uses a grid search algorithm for parameter search. Although the final solution is globally optimal, the training stage takes too long. As given in Table 5, the MA of grid search algorithm in algorithm B can reach 92%, $MASEN$ can reach 92.59%, and both MA_{mean} and MA_{median} are 92%, which also fully shows the global optimal performance and robustness of the grid search algorithm. Both the algorithm in this paper and algorithm A introduce the PSO algorithm, which is a local optimal algorithm. Therefore, the optimal parameter groups obtained after 20 searches are different, resulting in different test results.

As shown in Table 5, although the accuracy and sensitivity of the proposed algorithm are not all the best among them, the main purpose of this paper is to make both ACC and SEN achieve better results. According to the test result of the F-new function, as shown in Table 2, the maximum F-new function value F_{max} = 0.9357, and when the maximum value of F-new is achieved, ACC and SEN are 91% and 96.3%, respectively. The grid search algorithm obtains the global optimal solution, when ACC has a maximum value of 92%, corresponding to SEN = 92.59%, according to the F-new calculation method shown in equation (18), the F-new value is 0.9229, which is lower than the F-new value of 0.9357 obtained by the algorithm in this paper. Through

Table 4 AUC values of the four kernel functions

Kernel Function	RBF	Polynomial (d=2)	Polynomial (d=3)	Proposed
AUC	0.9498	0.9584	0.9584	0.9777

Table 5 Experimental statistics of introducing cost-sensitive and F-new function in the testing stage

	Parameter	Testing result							
Experiment	d	MA_{mean}	MA_{median}	MA	$MASEN$	MS_{mean}	MS_{median}	MS	$MSACC$
CS-MKL-SVM-PSO	3	88.95%	89.00%	92.00%	88.89%	93.52%	96.30%	96.30%	91.00%
algorithm A	3	90.40%	90.00%	91.00%	88.89%	86.67%	85.19%	88.89%	91.00%
algorithm B	2	92.00%	92.00%	92.00%	92.59%	92.59%	92.59%	92.59%	92.00%

the dual reference of cost-sensitive and F-new function, the model can achieve more than 90% of both ACC and SEN indexes under unbalanced data sets. From the perspective of the F-new function of the harmonic mean index of ACC and SEN, the CS-MKL-SVM-PSO algorithm proposed in this paper achieved the best performance.

In the experiment of the algorithm during the testing stage, from the perspective of ACC, MA_{mean} is reduced compared with algorithm A and algorithm B, but the obtained MA is the same as algorithm B at 92%, indicating that the algorithm in this paper still has the ability to seek better ACC. The overall recognition accuracy for ACC index is on the same level as the original algorithm. However, the SEN index of the nodule detection rate was considered in this paper. During the 20 experiments on the test set, the algorithm of this paper obtains the maximum sensitivity value MS of 96.3%, when the sensitivity is at the maximum, the accuracy $MSACC$ is 91%. Compared with that of algorithm A, the MS index of the algorithm in this paper is 7.41% higher and the $MSACC$ index is flat. It is observed that the MS index of the algorithm in this paper is 3.71% higher than that in algorithm B, while the $MSACC$ index is 1% lower than that in algorithm B. Therefore, the algorithm in this paper can improve the SEN index and reduce the missed detection of nodules to a certain extent.

In summary, the cost-sensitive algorithm proposed in this paper improves the detection of nodules by introducing different penalty coefficients to the positive and negative samples, and using the F-new function instead of the fitness function when the data set is seriously unbalanced. In terms of the time cost of the training algorithm model, the training time of the algorithm in this paper is slightly longer than that of the MKL-SVM-PSO algorithm (Li et al. 2018) but only 1/20 of the training time of the grid search algorithm (Li et al. 2013), which is easier to implement the online algorithm. Compared with using a single ACC or SEN as the objective function, using the F-new function can improve the detection rate of nodules. Therefore, the CS-MKL-SVM-PSO algorithm proposed in this paper has better recognition performance, which can take into account the overall recognition accuracy of nodules and reduce the missed detection of nodules.

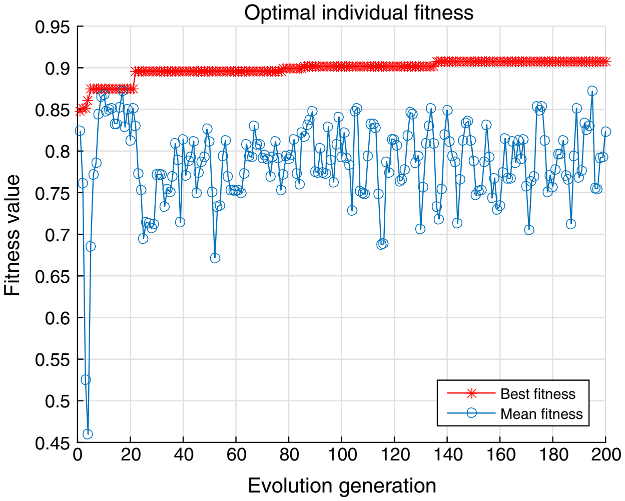


Fig. 5 Fitness curve of the CS-MKL-SVM-PSO algorithm

When the inertia weight is constant ω=1, the fitness curve of the algorithm in this paper is shown in Fig. 5.

As shown from Fig. 5, the fitness curve of the algorithm has obvious oscillation, and the convergence speed is slow. The algorithm performance can be further improved through a dynamic adjustment of parameters such as the inertia weight.

5 Conclusion

In this paper, first, a CS-MKL-SVM-PSO algorithm is proposed. Second, a new score function is proposed, and the MKL-SVM optimal parameter set can be found quickly and accurately. Finally, the proposed CS-MKL-SVM-PSO algorithm based on the score function as its objective function is applied to pulmonary nodule recognition. The main innovations of this work are as follows:

1. The cost-sensitive algorithm and MKL-SVM-PSO algorithm are combined to form the new CS-MKL-SVM-PSO algorithm. The cost-sensitive algorithm is used to solve the problem that positive and negative samples are imbalanced. However, the PSO algorithm is used to solve the problem of parameter optimization velocity, and the training model establishment can be accelerated.
2. A new F-new function is proposed to replace the fitness function by constructing the harmonic mean of ACC and SEN to consider both ACC and SEN and ultimately prevent the missed detection of nodules.
3. On the basis of constant weight, the experimental results of the CS-MKL-SVM-PSO algorithm, MKL-SVM-PSO algorithm and grid optimization MKL-SVM algorithm are analyzed.

The experimental results show that the proposed CS-MKL-SVM-PSO algorithm can effectively prevent the missed detection of nodules when the data set is unbalanced. Both ACC and SEN can be considered simultaneously with the proposed F-new function.

Although the above improvements have better experimental performance, there are still several problems and challenges to be solved in the future. In the future, the research work will focus on the following contents:

1. The PSO algorithm is a kind of local search algorithms, which can get the optimal solution faster, but its disadvantage is that with the increase in iterations, the diversity of the population decreases, which easily causes the phenomenon of premature particles, thus it is easy to fall into the local optimum. Therefore, some indexes are not reaching the optimum. The grid search algorithm can find the global optimal solution, but as the parameters increase, the training iteration period becomes longer. In order to find the optimal parameter set and reduce the training time, other global optimization algorithms and hybrid optimization algorithms will be explored to improve the model performance.
2. The kernel function directly determines the structure of the feature space, and the design of a new kernel function suitable for specific problems remains to be discussed.
3. At the same time, under the background of the rapid development of deep learning, how can the combination of traditional machine learning methods and deep learning be more helpful to promote the development of Lung CAD and better serve the medical field.

Acknowledgements The authors gratefully acknowledge the support of the National Natural Science Foundation of China (NSFC 61806024), the Foundation of Jilin Provincial Development of Science and Technology (20210201081GX, 20200401103GX) and Health Research Talents Special Project of Jilin Province (2020SCZ25).

Declarations

Conflict of interest The authors declare that they have no conflict of interest.

References

Abe S (2010) Support vector machines for pattern classification. Springer, Cham

Aoyama M, Li Q, Katsuragawa S, Li F, Sone S, Doi K (2003) Computerized scheme for determination of the likelihood measure of malignancy for pulmonary nodules on low-dose ct images. Med Phys, 30(3):387–394

Bansal G, Chamola V, Narang P, Kumar S, Raman S (2020) Deep3dscan: deep residual network and morphological descriptor based framework for lung cancer classification and 3d segmentation. IET Image Proc 14(7):1240

Bucak SS, Jin R, Jain AK (2014) Multiple kernel learning for visual object recognition: a review. IEEE Trans Pattern Anal Mach Intell 36(7):1354–1369

Cao W, Chen H, Yu Y, Li N, Chen W (2021) Changing profiles of cancer burden worldwide and in china: a secondary analysis of the global cancer statistics. Chin Med J 134(7):783–791

Chang CC, Lin CJ (2011) Libsvm: a library for support vector machines. ACM Trans Intell Syst Technol (TIST) 2(3):27

da Nóbrega RVM, Rebouças Filho PP, Rodrigues MB, Da Silva SPP, Dourado Júnior CMJM, De Albuquerque VHC (2018) Lung nodule malignancy classification in chest computed tomography images using transfer learning and convolutional neural networks. Neural Comput Appl

de Carvalho Filho AO, Silva AC, de Paiva AC, Nunes RA, Gattass M (2017) Computer-aided diagnosis of lung nodules in computed tomography by using phylogenetic diversity, genetic algorithm, and svm. J Digit Imaging 30(6):812–822

Demir Ö, Yılmaz Çamurcu A (2015) Computer-aided detection of lung nodules using outer surface features. Bio-Med Mater Eng 26(s1):S1213–S1222

Francisco Azuaje (2019) Artificial intelligence for precision oncology: beyond patient stratification. Npj Prec Oncol 3:1

Gao X, Chu C, Li Y, Lu P, Wang W, Liu W, Yu L (2015) The method and efficacy of support vector machine classifiers based on texture features and multi-resolution histogram from 18f-fdg pet-ct images for the evaluation of mediastinal lymph nodes in patients with lung cancer. Eur J Radiol 84(2):312–317

Gonçalves L, Novo J, Cunha A, Campilho A (2017) Learning lung nodule malignancy likelihood from radiologist annotations or diagnosis data. J Med Biol Eng 38(1):1–19

Guns R, Lioma C, Larsen B (2012) The tipping point: F-score as a function of the number of retrieved items. Inform Proc Manag 48(6):1171–1180

Hongtao X, Dongbao Y, Nannan S, Zhineng C, Yongdong Z (2018) Automated pulmonary nodule detection in ct images using deep convolutional neural networks. Pattern Recogn 85:109

Javaid M, Javid M, Rehman MZU, Shah SIA (2016) A novel approach to cad system for the detection of lung nodules in ct images. Comput Methods Prog Biomed 135:125–139

Jia D, Li Z, Zhang C (2020) Detection of cervical cancer cells based on strong feature cnn-svm network. Neurocomputing 411:112–127

Jk Liu, Hy Jiang, Cg He, Wang Y, Wang P, Ma H et al (2017) An assisted diagnosis system for detection of early pulmonary nodule in computed tomography images. J Med Syst 41(2):30

Jose D, Chithara AN, Kumar PN, Kareemulla H (2017) Automatic detection of lung cancer nodules in computerized tomography images. Nat Acad Sci Lett 40(3):161–166

Kauczor HU, Bonomo L, Gaga M, Nackaerts K, Peled N, Prokop M, Remy-Jardin M, von Stackelberg O, Sculier JP, of Radiology (ESR) ES, the European Respiratory Society (ERS) (2015) Esr/ers white paper on lung cancer screening. Eur Radiol 25(9):2519–2531

Li S, Xu P, Li B, Chen L, Wang J (2019) Predicting lung nodule malignancies by combining deep convolutional neural network and handcrafted features. Phys Med Biol 64(17):175012

Liao W, Wu J (2015) Matrix inequalities for the difference between arithmetic mean and harmonic mean. Ann Funct Anal 6(3):191–202

Lin CJ, Jeng SY, Chen MK (2020) Using 2d cnn with taguchi parametric optimization for lung cancer recognition from ct images. Appl Sci 10(7):2591

Li Y, Wen D, Wang K, et al. (2013) Mixed kernel function svm for pulmonary nodule recognition. In: International conference on image analysis and processing, Springer, pp 449–458

Li Y, Zhu Z, Hou A, Zhao Q, Liu L, Zhang L (2018) Pulmonary nodule recognition based on multiple kernel learning support vector machine-pso. Computational and mathematical methods in medicine 2018

Manikandan T, Bharathi N (2016) Lung cancer detection using fuzzy auto-seed cluster means morphological segmentation and svm classifier. J Med Syst 40(7):181

Monkam P, Qi S, Ma H, Gao W, Qian W (2019) Detection and classification of pulmonary nodules using convolutional neural networks: a survey. IEEE Access 7:78075

Orozco HM, Villegas OOV, Sánchez VGC, Domínguez HdJO, Alfaro MdJN (2015) Automated system for lung nodules classification based on wavelet feature descriptor and support vector machine. Biomed Eng Online 14(1):9

Parkash O, Thukral AK (2010) Statistical measures as measures of diversity. Int J Biomath 3(02):173–185

Polat H, Danaei Mehr H (2019) Classification of pulmonary ct images by using hybrid 3d-deep convolutional neural network architecture. Appl Sci 9(5):940

Puttagunta M, Ravi S (2021) Medical image analysis based on deep learning approach. Multimedia Tools Appl 13:1–34

Rijsbergen CJ (1974) Foundation of evaluation. J Doc 30(4):365–373

Siegel RL, Miller KD, Fuchs HE, Jemal A (2021) Cancer statistics, 2021. CA: Cancer J Clinic 71(1):7

Stewart B, Wild CP, et al. (2019) World cancer report 2014. International Agency for Research on Cancer

Sui Y, Wei Y, Zhao D (2015) Computer-aided lung nodule recognition by svm classifier based on combination of random undersampling and smote. Computational and mathematical methods in medicine 2015

Tajbakhsh N, Suzuki K (2016) Comparing two classes of end-to-end machine-learning models in lung nodule detection and classification: Mtanns vs cnns. Pattern Recogn 63:476

Tao X, Li Q, Guo W, Ren C, Li C, Liu R, Zou J (2019) Self-adaptive cost weights-based support vector machine cost-sensitive ensemble for imbalanced data classification. Inf Sci 487:31–56

Tong C, Liang B, Su Q, Yu M, Hu J, Bashir AK, Zheng Z (2021) Pulmonary nodule classification based on heterogeneous features learning. IEEE J Sel Areas Commun 39(2):574–581

Wang Q, Zhu W, Wang B (2015) Three-dimensional svm with latent variable: application for detection of lung lesions in ct images. J Med Syst 39(1):171

Wang T, Zhang L, Hu W (2021) Bridging deep and multiple kernel learning: a review. Inform Fusion 67(2):3–13

Ye Y, Tian M, Liu Q, Tai HM (2020) Pulmonary nodule detection using v-net and high-level descriptor based svm classifier. IEEE Access 8:176033–176041

Zhang B, Qi S, Monkam P, Li C, Yang F, Yao Y, Qian W (2019) Ensemble learners of multiple deep cnns for pulmonary nodules classification using ct images. IEEE Access 7:110358–110371

Zhao X, Liu L, Qi S, Teng Y, Wei Q (2018) Agile convolutional neural network for pulmonary nodule classification using ct images. Int J Comput Assist Radiol Surg 13(1):1–11

Zhou T, Lu H, Zhang J, Shi H (2016) Pulmonary nodule detection model based on svm and ct image feature-level fusion with rough sets. BioMed research international 2016