

Automatic detection of lung cancer nodules by employing intelligent fuzzy c-means and support vector machine.

Sakthivel K^{1*}, Jayanthiladevi A², Kavitha C³

¹Department of Computer Science and Engineering, K S Rangasamy College of Technology, Tamil Nadu, India

²Department of Computer Applications, Jain University, Bangalore, Karnataka, India

³Department of Computer Science, Thiruvalluvar Government Arts College, Tamil Nadu, India

Abstract

Segmentation is an essential step in image systems for the accurate lung disease diagnosis, since it delimits lung structures in Computerised Tomography (CT) images. Indeed, image processing techniques can help computer diagnosis if lung region is accurately obtained. A conventional fuzzy c-means clustering algorithm that has been implemented for segmentation of the Computerised Tomography (CT) lung images still suffers with low convergence rate, getting stuck in the local minima and vulnerable to initialization sensitivity. The proposed system presents an intelligent and dynamic approach called Intelligent Fuzzy C-Means (IFCM) to segment the lung nodules automatically and classify the lung nodules effectively using support vector machine classifier. This approach uses the capability of firefly search to find optimal initial cluster centers for the Fuzzy C-Means (FCM) and thus improve the segmentation accuracy. The features are extracted using fused tamura and haralick features after segmentation. These features are trained using different kernels of support vector machine for automatic detection of lung nodules as benign or malignant. The performance of support vector machine is evaluated by computing different measures from confusion matrix.

Keywords: Segmentation, Firefly search, Fuzzy c-means, Support vector machine, Confusion matrix.

Accepted on August 11, 2016

Introduction

Computer-Aided Design (CAD) systems for medical images typically involve the steps of segmentation of the image and classification of that segmented area. Various algorithms from different authors can be found for medical image segmentation such as thresholding, edge based and region based clustering. These methods may be effective for specific types of disease; segmentation of lungs is always a challenging problem due to changes in pathology in the parenchyma area, or in shape and anatomic connection to neighbouring pulmonary structures, such as blood vessels or pleura.

It could be a great clinical importance if patient's chance of survival is increased by an effective Computer-Aided Design (CAD) system [1]. The systems are developed to support the process of distinguishing malignant with benign lesions. The systems need to improve both the sensitivity and specificity in order to be successful for clinical use. Computer-Aided Design (CAD) methods may use features extracted either by the computer or by the radiologist [2]. For interpreting instantaneous data, radiologists employ many radiographic images and computer analysis is needed for determining the significant features.

Fuzzy c-means algorithm is popular algorithm for robust lung image segmentation. The fuzzy c-means algorithm attempts to partition a finite collection of n elements into a collection of c fuzzy clusters with respect to some given criterion. Many metaheuristic search algorithms have been hybridized with the fuzzy c-means algorithm to find optimal cluster centers. These algorithms explore the entire search space in the problem to determine possible solutions. These algorithms include bee optimization, harmony search, the ant colony algorithm, simulated annealing, the genetic algorithm, tabu search, the firefly algorithm and particle swarm. In this research work the data is in the form of Computer Tomography (CT) images of the lung and hence image analysis is done to perform diagnosis. The segmentation of such image is done by using an intelligent and dynamic approach called Intelligent Fuzzy C-Means with firefly search (IFCM) and classified effectively using support vector machine.

Related works

Image processing plays a vital role in diagnosis of diseases using automatic detection of lung nodules based on images. This research focuses on Computer-Aided Design (CAD) systems for diagnosis of lung cancer. As with any system

employing image analysis and image understanding, segmentation is one of the primary processes in a Computer-Aided Design (CAD) system that works with images. Segmentation of lung cancer is a challenging task [3]. The Region Of Interest (ROIs) can be extracted only if the lungs are correctly segmented [4]. If the segmentation is incorrect, it may increase the false positives and false negatives thereby affecting the diagnostic accuracy of the system. Hence, it is important to use a robust and reliable technique for segmentation.

The use of Artificial Intelligence (AI) in diagnostic expert systems aims at making the systems to mimic the decision-making process of human experts. However, such systems can be used only to provide a “second opinion” to the physicians in improving diagnostic accuracy and cannot be considered as a replacement for a physician. Artificial Intelligence (AI) techniques that are most frequently used in Computer-Aided Design (CAD) systems in the literature are Fuzzy logic, Genetic Algorithms, Neural Networks and ensemble approaches.

Image segmentation is more crucial part in obtaining information from the medical images. It is used for detecting the threatening diseases in medical field which lead to death. The lungs segmentation is a challenging problem due to the need of different scanners and scanning protocols for scanning the homogeneities in the lung region, pulmonary structures of similar densities such as arteries, veins, bronchi, and bronchioles [5]. The segmentation of lung regions from Computer Tomography (CT) images and chest radiographs by a particular technique can be addressed in terms of accuracy, processing time, and automation level. The major classification of lung segmentation techniques [6] are divided in to four categories: methods based on signal thresholding, deformable boundaries, shape models, edge based and region based.

Among all the segmentation methodologies mentioned in the literature, fuzzy c-means produce a robust segmentation results. But, the initial cluster centroid values and number of clusters should be given as input for fuzzy c-means. This limitation prevents traditional fuzzy c-means to become more popular in image segmentation. Initial cluster center and number of cluster could be dynamically generated using a popular metaheuristic search algorithm called firefly search algorithm. The hybridization of Intelligent Fuzzy C-Means with firefly search (IFCM) for segmenting the lung Computer Tomography (CT) images produces robust segmentation. Support vector machine classifier is superior among all other classifiers mentioned in the literature [7]. Hence, this research uses support vector machine for effective classification of lung nodules.

Materials and Methods

The proposed scenario considered the automatic detection of lung cancer nodules and classification of lung cancer nodules which is acquired from computed tomography. The data is collected from the histology banks and samples are sorted

across thousands of patients. These image data is pre-processed to extract the lung region alone. The Region Of Interest (ROI) is determined by using segmentation algorithm and features are extracted for classifying the cancerous or non-cancerous regions.

The proposed computer aided diagnosis systems comprises of four phases for early detection of lung cancer and are as follows:

Phase 1: Image denoising.

Phase 2: Segmentation of lung region and Region Of Interest (ROIs).

Phase 3: Feature extraction from the segmented region.

Phase 4: Classification of malignant and benign nodules.

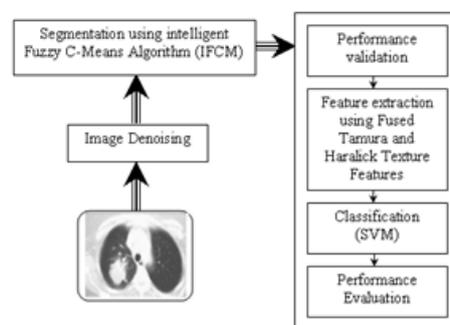


Figure 1. Schematic flow of the proposed system.

Figure 1 show the proposed system implemented with all the phases of Computer-Aided Design (CAD) system. In the initial phase of this Computer-Aided Design (CAD) method, Computer Tomography (CT) lung image is denoised using median filter. Next image is segmented using Intelligent Fuzzy C-Means (IFCM) and segmentation quality is validated. After segmentation tamura and haralick texture features are extracted to classify benign or malignant using support vector machine which automatically detect lung cancer nodules [8]. Finally the performance is evaluated using confusion matrix.

Segmentation by proposed Intelligent Fuzzy C-means (IFCM) algorithm: fuzzy clustering is the process of assigning the membership value and using them to assign elements to one or more clusters [9]. The commonly used fuzzy method for clustering is Fuzzy C-Means (FCM) [10]. The aim of the objective function of Fuzzy C-Means (FCM) is to find the cluster centers and to produce class membership matrix, which designates a membership to a data point, depending on the relativity of the data point to a particular class when compared to other classes. The performance of the firefly algorithm in terms of obtaining near-optimal cluster centers values in the initialization phase in the Fuzzy C-Means (FCM) algorithm is investigated [11]. A clustering approach based on Fuzzy C-Means is proposed [12]. The proposed clustering method consists of two phases:

In order to determine the optimal cluster centers, firefly inspects the search space of the given dataset and then the values of the cluster centers will be obtained using the firefly

algorithm. The pseudo code for the proposed Intelligent Fuzzy C-Means (IFCM) algorithm is given below and obtained result is shown in Figure 2.

Phase I

Begin

Initialize the parameters of the firefly algorithms:

- Fireflies number (n).
- Maximum number of generations (Max-generation, iterations).
- β_0 , α and γ

Generate the initial population (n initial solutions) of fireflies

$X_i = (i=1, 2, 3, \dots, n)$

The intensity of the light at will be determined using the objective function value $F(x)$.

Determine the absorption (assimilation) coefficient γ .

While ($m < \text{Maxgeneration}$)

For $i=1: n/n$ -number of all fireflies

For $j=1: n/n$ -number of all fireflies

If ($I_j > I_i$)

Move firefly i towards j in d -dimension

End if

Get attractiveness, which differs with distance r via $\exp[-\gamma r]$

Calculate the new solutions and update light intensity

End for j

End for i

Phase II

Initialize fuzzy c-means cluster center by new solution of firefly algorithm

For $t = \text{to max Iterations do}$

Update the membership matrix u_{ij}

Calculate the new cluster centers v_{ij}

Calculate the new objective function J_m

If ($\text{abs}(U^{(k+1)} - U^{(k)}) < \epsilon$) then

Break;

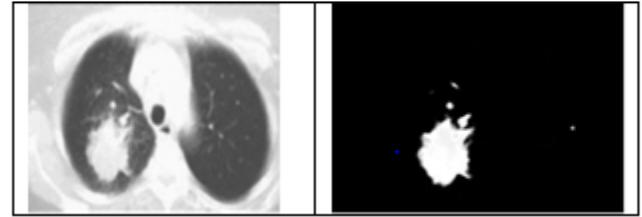
Else

$U^{(k+1)} = U^{(k)}$

End if

End for

End



a) Original image

b) Tumour segmentation

Figure 2. Results obtained for computer tomography lung image using proposed intelligent fuzzy c-means with firefly search.

Segmentation accuracy

The segmentation accuracy A_s [13] is calculated using Equation 1 to evaluate the quality of the segmented lung computer tomography image using Intelligent Fuzzy C-Means (IFCM) with firefly search

$$A_s = \frac{S_{prop}}{O_{prop}} \times 100 \rightarrow (1)$$

Where S_{prop} and O_{prop} is defined as

$S_{prop} = (\text{Solidity} + \text{Area} + \text{Perimeter})$ segmented

$O_{prop} = (\text{Solidity} + \text{Area} + \text{Perimeter})$ unsegmented

The segmentation accuracy A_s is computed and the results are shown in Table 1 and it is observed that proposed Intelligent Fuzzy C-Means (IFCM) performs better segmentation.

Table 1. Segmentation accuracy for Fuzzy Possibilistic C- Means (FPCM) and Proposed Intelligent Fuzzy C-Means (IFCM).

Segmentation methodologies	Accuracy (%)
Fuzzy Possibilistic C- Means (FPCM)	82.67
Proposed Intelligent Fuzzy C-Means (IFCM)	96.63

Automatic detection of lung cancer nodules using support vector machine

Texture features are important which is used for identifying objects or Regions Of Interest (ROIs) in a lung computer tomography image. Haralick and Tamura texture features are extracted from the segmented region. Both features will be fused to train the binary support vector machine classifier with the assumption that label 1 belongs to malignant and label 0 belongs to benign. Three kernels such as linear kernel, polynomial kernel and radial basis function are used to construct support vector machine classifier [14]. The most used kernel function for support vector machine is Radial Basis Function (RBF) because of their localized and finite responses across the entire range of real x-axis. The classification accuracy of Radial Basis Function (RBF) kernel was high; also, the bias value and the error rate of Radial Basis Function (RBF) kernel were small when compared to other kernels. After the learning process is completed by providing several

conditions, the proposed technique would be able to detect the presence of cancer in the lung region automatically.

Performance Evaluation

A total of 400 images are considered in this work of which 200 belong to benign and the rest belong to malignant cases. The input datasets are grouped into two; a training set and a testing set with 200 datasets each. The performance of the proposed Computer-Aided Design (CAD) system is evaluated using the following parameters based upon the confusion matrix as shown in Table 2.

Accuracy (ACC)

Sensitivity (SEN)

Specificity (SPE)

Positive Predictive Value (PPV)

Negative Predictive Value (NPV)

F-Score (FS)

Table 2. Confusion matrix.

98 (TP)	3 (FP)
02 (FN)	97 (TN)

TP: True Positives; FP: False Positives; FN: False Negatives; TN: True Negatives.

Accuracy is a measure that rates the overall effectiveness of the system. Referring to Table 3, the sensitivity and specificity of Radial Basis Function (RBF) kernel are 98% and 97% respectively. The false positive proportion using Radial Basis Function (RBF) kernel is 3% and the false negative proportion using Radial Basis Function (RBF) kernel is 2%. False positive and false negative proportions are the complements of sensitivity and specificity respectively.

Table 3. Performance measures obtained using different kernels of support vector machine.

Measures	Linear (%)	Polynomial (%)	RBF (%)
Sensitivity	92	94	98
Specificity	87	93	97
Accuracy	89.5	93.5	97.6
PPV	87.62	93.07	97.03
NPV	91.58	93.94	97.98
F-Score	89.75	93.53	97.51

PPV: Positive Predictive Value; NPV: Negative Predictive Value; RBF: Radial Basis Function.

Table 4. Performance comparison of proposed system with existing computer-aided design systems.

Computer-Aided Design Systems	Sensitivity (%)	Specificity (%)	Accuracy (%)	FP rate (%)
Mabrouk et al. [13]	94.1	95	94.9	0.05
Ganesh et al. [15]	96.15	94.13	95.56	0.05
Proposed Intelligent Fuzzy C-Means (IFCM)	98	97	97.6	0.03

The positive predictive value of radial basis function kernel is 97.03% which indicates the proportion of the patients with positive test results who are correctly diagnosed and it is higher compared to the other methods. The F-score is a measure of a test's accuracy. The F-Score can be interpreted as a weighted average of the precision and recall and takes its best value as 1 and worst score as 0. The F-score of radial basis function kernel is closer to 1. The accuracy obtained using the radial basis function kernel is 97.6% which is higher than that those obtained using other kernels. Table 3 shows the overall performance measures of linear, polynomial and radial basis function kernels of support vector machine. From this, it is inferred that support vector machine radial basis function kernel outperforms the other two kernels.

Automated Computer-Aided Design (CAD) systems are useful for radiologists for detection and diagnose of lung cancer from computer tomography images. Table 4 shows the comparative results of various existing Computer-Aided Design (CAD) systems [13,15] with the proposed system. The proposed system outperforms the state-of-art Computer-Aided Design (CAD) systems with the accuracy of 97.6% and the average false positive ratio of 0.03%. It is clear that, when sensitivity and specificity increases, the presence of the malignant and benign nodules are classified correctly. The accuracy of the Computer-Aided Design (CAD) system will also improve. If the sensitivity and specificity decreases, the accuracy of the Computer-Aided Design (CAD) system will decrease which leads to misclassification of malignant and benign nodules. From the observation, it is proved that the proposed Computer-Aided Design (CAD) system performs better when compared with the other state-of-art Computer-Aided Design (CAD) systems.

Conclusion

In this study, intelligent fuzzy-c means algorithm have been presented for segmenting the computer tomography lung nodules. Initially the images are denoised and applied with the proposed segmentation method. The segmentation accuracy is better when compared with the Fuzzy Probabilistic C-Means (FPCM). A support vector machine classifier is constructed and trained using fused texture features which are extracted from segmented region. The performances are evaluated and the accuracy achieved is 97.6% with the average false positive ratio of 0.03%. From the observation, it is proved that the proposed Computer-Aided Design (CAD) system employing intelligent fuzzy-c means and support vector machine performs

better when compared with the other state-of-art Computer-Aided Design (CAD) systems.

References

1. Shaik PS, Kavitha C. A review on computer aided detection and diagnosis of lung cancer nodules. *Int J Comp Technol* 2012; 3: 393-400.
2. Sruthi I, Robin J. Computer aided lung cancer detection system. *Proc Glob Conf Commun Technol* 2015; 555-558.
3. Hu S, Hoffman EA, Reinhardt JM. Automatic lung segmentation for accurate quantitation of volumetric X-ray CT images. *IEEE Trans Med Imaging* 2001; 20: 490-498.
4. Sluimer I, Prokop M, van Ginneken B. Toward automated segmentation of the pathological lung in CT. *IEEE Trans Med Imaging* 2005; 24: 1025-1038.
5. Punithavathy KR, Sumathi PMM. Analysis of statistical texture features for automatic lung cancer detection in PET/CT images. *Robo Autom Contr Embed Sys* 2015; 1-5.
6. Chonglun L. A new automatic seeded region growing algorithm. *Proceedings of 6th International Congress on Image and Signal Processing (CISP)* 2013; 543-549.
7. Gomathi M, Thangaraj P. A computer aided diagnosis system for lung cancer detection using machine learning technique. *Europ J Sci Res* 2011; 51: 260-275.
8. Jiangdian S, Caiyun Y, Li F, Kun W, Feng Y, Shiyuan L, Jie T. Lung lesion extraction using a toboggan based growing automatic segmentation approach. *IEEE Trans Med Imaging* 2016; 35: 337-353.
9. Venkatasalam K, Rajendran P. Effective RBIR Fuzzy C-Means segmentation HAAR wavelet with user interactive multi threshold robust features vector. *Asian J Inform Technol* 2016; 15: 223-231.
10. Gong M, Liang Y, Shi J, Ma W, Ma J. Fuzzy C-means clustering with local information and kernel metric for image segmentation. *IEEE Trans Image Process* 2013; 22: 573-584.
11. Saeid N, Amin AN, Saeid H, Abdolreza S, Farhad S. Optimization of KFCM clustering of hyperspectral data by particle swarm optimization algorithm. *Intl J Humanities* 2013; 20: 101-120.
12. Alomoush WS, Abdullah SS, Hussain R. Segmentation of MRI brain images using FCM improved by firefly algorithms. *J Applied Sci* 2013; 14: 66.
13. Faizal Khan Z, Kannan A. Intelligent approach for segmenting Computer tomography lung images using fuzzy logic with bitplane. *J Electr Eng Technol* 2014; 9: 742-750
14. Mai M, Ayat K, Amr S. Support vector machine based computer aided diagnosis system for large lung nodules classification. *J Med Imaging H Inform* 2013; 3: 214-220.
15. Ganesh S, Harsha B. Lung segmentation and tumor identification from computer tomography scan images using support vector machine. *Int J Sci Res* 2014; 3.

*Correspondence to

Sakthivel K

Department of Computer Science and Engineering

K S Rangasamy College of Technology

Tamil Nadu

India