Lung Nodule Detection Using Fuzzy Clustering and Support Vector Machines

S.Sivakumar [#1], Dr.C.Chandrasekar [*2]
[#] Department of Computer Science
Periyar University, Salem-636 011, India
[1] ssivakkumarr@yahoo.com
[*] Department of Computer Science
Periyar University, Salem-636 011, India
[2] ccsekar@gmail.com

Abstract— **Lung cancer is the primary cause of tumor deaths for both sexes in most countries. Lung nodule, an abnormality which leads to lung cancer is detected by various medical imaging techniques like X-ray, Computerized Tomography (CT), etc. Detection of lung nodules is a challenging task since the nodules are commonly attached to the blood vessels. Many studies have shown that early diagnosis is the most efficient way to cure this disease. This paper aims to develop an efficient lung nodule detection scheme by performing nodule segmentation through fuzzy based clustering models; classification by using a machine learning technique called Support Vector Machine (SVM). This methodology uses three different types of kernels among these RBF kernel gives better class performance.**

Keywords-Lung cancer, Image segmentation, FCM, WPFCM, Classification, Support Vector Machines

I. INTRODUCTION

A. Lung Cancer:

Lung cancer is the primary cause of tumor deaths for both sexes in most countries. There are four stages of lung cancer from I to IV with rising gravity. If the cancer is detected at stage I and it has no more 30 mm in diameter, then there is about 67% survival rate, and only less than 1% chance left for stage IV. Thus it is concluded that early detection and treatment at stage 1 have high survival rate. But unfortunately, lung cancer is usually detected late due to the lack of symptoms in its early stages. This is the reason why lung screening programs have been investigated to detect pulmonary nodules: they are small lesions which can be calcified or not, almost spherical in shape or with irregular borders. The nodule definition for thoracic CT of the Fleischer's Society is "a round opacity, at least moderately well margined and no greater than 3 cm in maximum diameter" [7]. Approximately 40% of lung nodules are malignant, that is, are cancerous: the rest is usually associated with infections. Because malignancy depends on many factors, such as patient age, nodule shape, doubling time, presence of calcification [8], after the initial nodule detection further exams are necessary to obtain a diagnosis. In computer vision, segmentation refers to the process of partitioning a digital image into multiple regions or sets of pixels. Each of the pixels in a region is similar with respect to some characteristic or computed property, such as color, intensity, or texture. Adjacent regions are significantly different with respect to the same characteristics [9][10][11]. Early diagnosis has an important prognostic values and has a huge impact on treatment planning [1]. As nodules are the most common sign of lung cancer, nodule detection in CT scan images is a main diagnostic problem. Conventional projection radiography is a simple, cheap, and widely used clinical test. Unfortunately, its capability to detect lung cancer in its early stages is limited by several factors, both technical and observer-dependent. Lesions are relatively small and usually contrast poorly with respect to anatomical structure. This partially explains why radiologists are commonly credited with low sensitivity in nodule detection, ranging from 60 to 70%. A thorough review of the drawbacks affecting conventional chest radiography is given, for example, by Woodring [2]. However, several long-term studies carried out in the 1980s using large clinical data sets have shown that up to 90% of nodules may be correctly perceived retrospectively [3], [4]. In addition, detection sensitivity can be increased to more than 80% in the case of a double radiograph reading by two radiologists. Furthermore, sensitivity is expected to increase with the widespread use of digital radiography systems which are characterized by an extended dynamic range and have a better contrast resolution than conventional film radiography. In view of this, the availability of efficient and effective computer-aided diagnosis (CAD) systems is highly desirable [5], as such systems are usually conceived to provide the physician with a second opinion [6] so as to focus his/her attention on suspicious image zones, playing the role of a "second reader."

B. Image Segmentation:

Image segmentation is a necessary task for image understanding and analysis. A large variety of methods have been proposed in the literature. Image segmentation can be defined as a classification problem where each pixel is assigned to a precise class. Image segmentation is a significant process for successive image analysis tasks. In general, a segmentation problem involves the division a given image into a number of homogeneous segments, such that the union of any two neighboring segments yields a heterogeneous segment. Numerous segmentation techniques have been proposed earlier in literature. Some of them are histogram based technique, edge based techniques, region based techniques, hybrid methods which combine both the edge based and region based methods together, and so on. In recent years image segmentation has been extensively applied in medical field for diagnosing the diseases. Image segmentation plays an important role in a variety of applications such as robot vision, object recognition, and medical imaging [12]. In the field of medical diagnosis an extensive diversity of imaging techniques is presently available, such as radiography, computed tomography (CT) and magnetic resonance imaging (MRI). In recent times, Computed Tomography (CT) is the most effectively used for diagnostic imaging examination for chest diseases such as lung cancer, tuberculosis, pneumonia and pulmonary emphysema. The volume and the size of the medical images are progressively increasing day by day. Therefore it becomes necessary to use computers in facilitating the processing and analyzing of those medical images. Even though the original FCM algorithm yields good results for segmenting noise free images, it fails to segment images corrupted by noise, outliers and other imaging artifact. Medical image segmentation is an indispensable step for most successive image analysis tasks. This paper presents an image segmentation approach using standard Fuzzy C-Means (FCM), Fuzzy-Possibilistic C-Means and weighted Fuzzy-Possibilistic C-Means (WFCM) algorithm on the CT lungs images. From the segmented image features were extracted and these features are used as input for SVM classifier.

C. Support Vector Machines:

SVM is a machine learning tool, based on the idea of data classification. It performs classification by constructing an N-dimensional hyper plane that optimally separates the data into two categories. The separation of data can be either linear or non-linear. Kernel function maps the training data into a kernel space and the default kernel function is the dot product. For non-linear cases, SVM uses a kernel function which maps the given data into a different space; the separations can be made even with very complex boundaries. The different types of kernel function include polynomial, RBF, quadratic, Multi-Layer Perceptron (MLP). Each kernel is formulated by its own parameters like γ, σ, etc. By varying the parameters the performance rate of the SVM can be measured.

D. LIDC Dataset:

The Lung Image Database Consortium image collection (LIDC-IDRI) consists of diagnostic and lung cancer screening thoracic CT scans with marked-up annotated lesions. It is a web-accessible international resource for development, training, and evaluation of computer-assisted diagnostic (CAD) methods for lung cancer detection and diagnosis. The LIDC-IDRI collection contained on The Cancer Imaging Archive (TCIA) is the complete data set of all 1,010 patients which includes all 399 pilot CT cases plus the additional 611 patient CTs and all 290 corresponding chest x-rays. The lungs image data, nodule size list and annotated XML file documentations can be downloaded from the National Cancer Institute website:

https://wiki.cancerimagingarchive.net/display/Public/LIDC-IDRI.

This paper is organized into three different sections including this introduction to lung cancer and the discussion of the existing techniques. In section II the technique for segmentation of the lung nodules is proposed, which includes the FCM, FPCM and WFPCM. Section III discuss about SVM, its classification methodology using different types of kernels. Conclusion and future scopes are drawn in section IV.

II. PROPOSED METHOD

The proposed method of lung nodule detection is shown in Figure 1. All the lungs CT scan images are in the format of DICOM with the size of 512X512. To enhance the CT scan images, 3X3 window based median filter was applied to remove the noise. After enhancing the CT scan image the WFPCM algorithm is applied to segment the image. From the segmented image mean, standard deviation, contrast and entropy features are calculated.

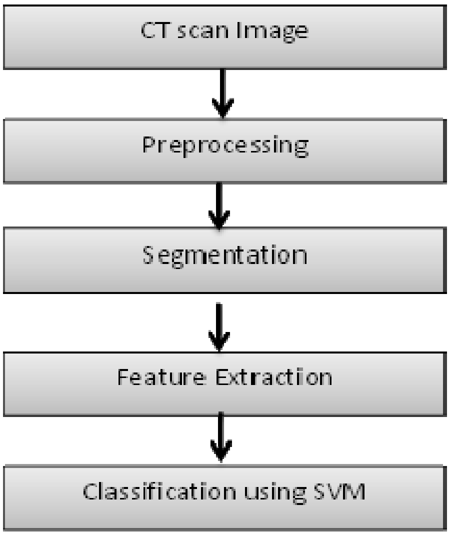


Fig.1. Block diagram of proposed method

A. *Fuzzy C-Means Clustering:*

In hard clustering, data is divided into distinct clusters, where each data element belongs to exactly one cluster. In fuzzy clustering, data elements can belong to more than one cluster, and associated with each element is a set of membership levels. These indicate the strength of the association between that data element and a particular cluster. Fuzzy clustering is a process of assigning these membership levels, and then using them to assign data elements to one or more clusters. One of the most widely used fuzzy clustering algorithms is the Fuzzy C-Means (FCM) Algorithm [14]. The FCM algorithm attempts to partition a finite collection of n elements $X = \{x_1,...,x_n\}$ into a collection of c fuzzy clusters with respect to some given criterion[13]. Given a finite set of data, the algorithm returns a list of c cluster centers $C = \{c_1,...,c_c\}$ and a partition matrix $U=u_{i,j} \in [0,1]$, $i=1,2,\ldots.n$, $j=1,\ldots.c$. where each element u_{ij} tells the degree to which element x_i belongs to cluster c_j. The standard FCM algorithm minimizing the following objective function,

$$J_m(U,V)=\sum_{i=1}^{c}\sum_{k=1}^{n}u_{ik}^{m}\left\|x_k-v_i\right\|^2, \qquad 1\leq \mathrm{m} < \infty \qquad (1)$$

Where c is the number of clusters, n *is the* number of data points, u_{ik} is the membership of x_k in class i, satisfying $\sum_{ik}^{c} u_{ik} = 1$, where m is the fuzziness value and v is the set of cluster centers.

The FCM algorithm consists of the following steps:

1. *Initialize* $U=[u_{ij}]$ *matrix,* $U^{(0)}$
2. *At k-step: calculate the centers vectors* $C^{(k)}=[c_j]$ *with* $U^{(k)}$ using (2)

$$c_j = \frac{\sum_{i=1}^{N} u_{ij}^{m}.x_i}{\sum_{i=1}^{N} u_{ij}^{m}} \qquad (2)$$

3. *Update* $U^{(k)}$, $U^{(k+1)}$ using (3)

$$u_{ij} = \frac{1}{\sum_{k=1}^{c}\left(\frac{\left\|x_i-c_j\right\|}{\left\|x_i-c_k\right\|}\right)^{\frac{2}{m-1}}} \qquad (3)$$

4. *If* $\| U^{(k+1)} - U^{(k)} \| < \varepsilon$ *then STOP; otherwise return to step 2.*

B. *Fuzzy-Possibilistic C-Means:*

FPCM algorithm was proposed by N.R.Pal, K.Pal, and J.C.Bezdek [14] and it includes both possibility and membership values. FPCM model can be seen as below:

$$min\{J_{M,\eta}(U,T,V;X)\} = \sum_{i=1}^{c}\sum_{k=1}^{n}\left(u_{ik}^{m} + t_{ik}^{\eta}\right)\ \left\|x_j - a_i\right\|^2 \qquad (4)$$

Where U is membership matrix, T is possibilistic matrix, and V is the resultant cluster centers, c and n are cluster number and data point number respectively.

$$\mu_{ij} = \frac{1}{\sum_{j=1}^{c}\left(\frac{\left\|x_j - a_i\right\|^2}{\left\|x_k - a_j\right\|^2}\right)^{\frac{2}{m-1}}}, 1 \leq i \leq\leq c; 1 \leq k \leq n \qquad (5)$$

$$t_{ik} = \frac{1}{\sum_{j=1}^{c}\left(\frac{\left\|x_j - a_i\right\|^2}{\left\|x_k - a_j\right\|^2}\right)^{\frac{2}{\eta-1}}}, 1 \leq i \leq c; 1 \leq k \leq n \qquad (6)$$

$$v_i = \frac{\sum_{k=1}^{n}(u_{ik}^{m}+t_{ik}^{n})x_k}{\sum_{k=1}^{n}(u_{ik}^{m}+t_{ik}^{n})}, 1 \leq i \leq c \qquad (7)$$

The above equations show that membership u_{ik} is affected by all c cluster centers, while possibility t_{ik} is affected only by the *i*-th cluster center *c*. The possibilistic term distributes the t_{ik} with respect to all *n* data points, but not with respect to all *c* clusters. So, membership can be called relative typicality, it measures the degree to which a point belongs to one cluster relative to other clusters and is used to crisply label a data point. And possibility can be viewed as absolute typicality, it measures the degree to which a point belongs to one cluster relative to all other data points, it can reduce the effect of outliers. Combining both membership and possibility can lead to better clustering result.

C. *Weighted Fuzzy-Possibilistic C-Means:*

The objective function of the Weighted Fuzzy-Possibilistic Clustering can be formulated as follows[17]:

$$J_{WFPCM} = \sum_{i=1}^{c}\sum_{j=1}^{n}\left(\mu_{ij}^{2m}w_{ji}^{m}\left\|x_j - a_i\right\|^{2m} \quad + t_{ij}^{2\eta}w_{ji}^{m}\left\|x_j - a_i\right\|^{2m}\right) \qquad (8)$$

Where

$$\mu_{ij} = \frac{1}{\sum_{k=1}^{c}\left(\frac{\left\|x_j - a_i\right\|^{2m}}{\left\|x_k - a_j\right\|^2}\right)^{\frac{2m}{m-1}}} \qquad (9)$$

$$t_{ij} = \frac{1}{\sum_{k=1}^{c}\left(\frac{\left\|x_j - a_i\right\|^{2m}}{\left\|x_k - a_j\right\|^2}\right)^{\frac{2\eta}{\eta-1}}} \qquad (10)$$

$$v_i = \frac{\sum_{j=1}^{n}\left(\mu_{ij}^{2m}w_{ji}^{2m}+t_{ij}^{2\eta}w_{ji}^{2\eta}\right).x_j}{\sum_{j=1}^{n}\left(\mu_{ij}^{2m}w_{ji}^{2m}+t_{ij}^{2\eta}w_{ji}^{2\eta}\right)} \qquad (11)$$

$$w_k = \sum_{y=1}^{n} \exp\left(-h \times \left\|x_k - x_y\right\| / \sigma\right) \qquad (12)$$

where h is a resolution parameter and σ = standard deviation of input data. The clustered data will be validated against the following indices [15]: Partition Coefficient (PC), Classification Entropy (CE), Partition Index (SC), Separation Index(S), and Xie-Beni Index (XB) which shows WFPCM produces better partition.

D. *Feature Extraction:*

The features extracted from the lung images are mean, contrast, entropy and standard deviation. Mean value denotes the average value of all the pixels. Contrast is a measure of the intensity between a pixel and its neighborhood of the image.

$$contrast = \sum_{i,j} |i - j|^2 p(i, j) \tag{13}$$

where i,j denotes the row and the column pixel.

$$standard\ deviation = \sqrt{\frac{1}{N} \sum_{I=1}^{N} (x_i - \bar{x})^2} \tag{14}$$

where, $\bar{x}$ is the mean value and n is the number of elements in the sample.

Entropy is a statistical measure of randomness, used to characterize the textural properties of input image. It is given by,

$$Entropy = (p .* \log(p)) \tag{15}$$

where “p” is the input image.

III. CLASSIFICATION USING SVM

A. *Support Vector Machines:*

SVM is a machine learning technique which is used as a classification tool. It uses kernel function, which acts upon the input data; final summation with an activation function gives the final classification result. The architecture of SVM is shown in Fig.2, in which the suffix “n” represents number of vectors. Ns denote the number of support vectors. A binary classification [16] is used here, in which a hyper plane classifies the given data into two different classes; the vectors closest to the boundaries are called support vectors and the distance between the support vectors and hyper plane is called margin.

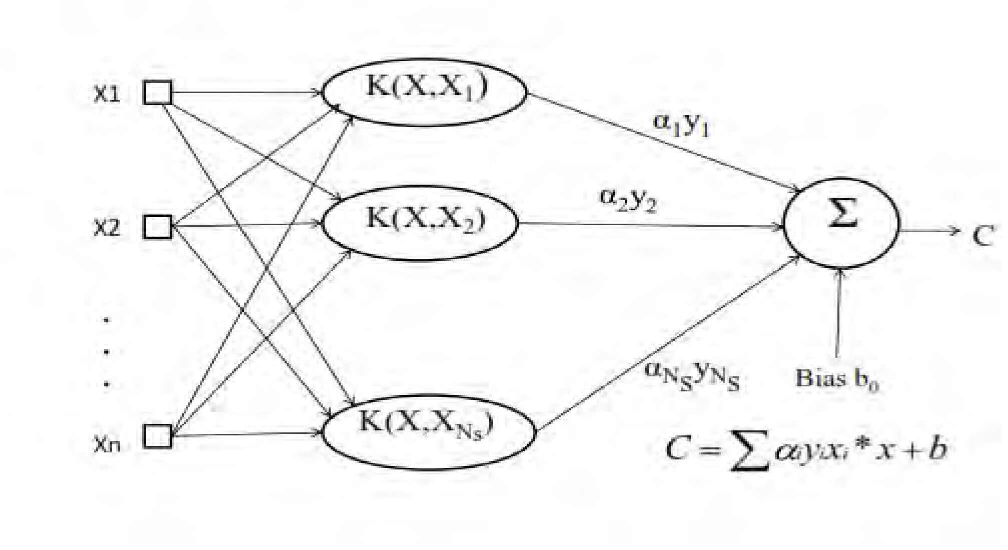


Fig.2 Architecture of SVM

SVM uses different types of kernels like linear, polynomial and RBF. Their formulations are given in the equations 13-15, in which the term “K(Xi,Xj)” represents the kernel function; Xi and Xj are the vectors under classification.

$$(Xi, Xj) = Xi^T Xj \tag{16}$$

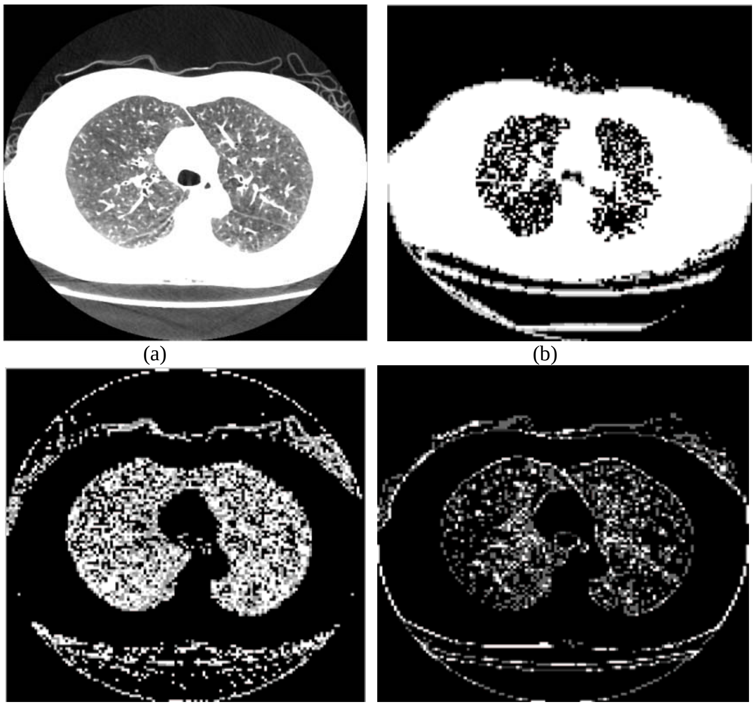
$$(Xi, Xj) = (\gamma Xi^T Xj + r) \tag{17}$$

$$(Xi, Xj) = ex(-\gamma\ Xi - Xj\ )^2 \tag{18}$$

In the above kernel function types, ‘γ’ and ‘d’ are the kernel parameters, whose values are 1 and 3 respectively. Totally, 54 lung images are taken from the LIDC database which includes 18 non-cancerous and 36 cancerous types.

B. *Result and Discussion:*

For the experiment we taken CT scan lung cancer affected images of size 512X512 with number of cluster is 5, with fuzziness value m=2 and e=0.0001. Fig.3. shows the segmentation results of FCM, FPCM and WFPCM.



(a) (b)

(c) (d)

Fig. 3. (a) Preprocessed CT scan lungs image (b) FCM applied image (c) FPCM applied image (d) WFPCM applied image

The segmented data validated against the Partition Coefficient (PC), Classification Entropy (CE), Partition Index (SC), Separation Index(S), and Xie-Beni Index (XB), which shows WFPCM, produces better partition compare with FCM and FPCM. The features extracted from the images are mean, contrast, entropy and standard deviation used as the input for SVM classifier. Table 1, shows the accuracy, specificity and sensitivity for the three different kernels of SVM classifier.

TABLE I
Result for Three Different Kernels of SVM Classifier

Kernel type	**Accuracy**	**Specificity**	**Sensitivity**
Linear	71.43%	75%	70.83%
Polynomial	66.07%	52.94%	71.79%
RBF	80.36%	76.47%	82.05%

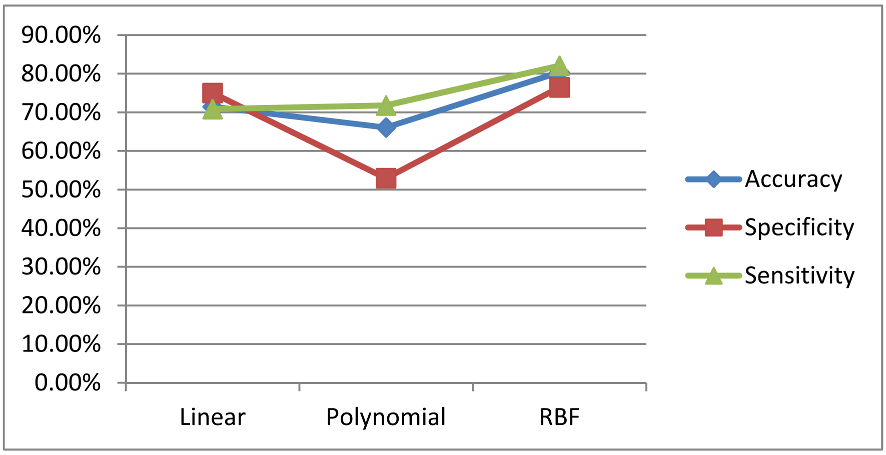


Fig.3. Performance analysis chart for SVM Classifier

From the Table 1 and Figure 3, the RBF kernel based SVM classifier performs better than linear and polynomial kernel based classifier.

IV. Conclusion

In this paper, a segmentation technique based on weighted fuzzy possibilistic based clustering is carried out for lung cancer images. Then classification of lung nodules as normal/abnormal is done by using SVM. In this paper, it is shown that RBF kernel gives better classification performance where compare with Linear and Polynomial kernels. The future work is to do the classification performance by using any other classifier.

References

[1] C. Society, Cancer Facts and Figures 2001. Atlanta, GA: American Cancer Society, 2001.

[2] J. Woodring.: “Pitfalls in the radiologic diagnosis of lung cancer,” AJR, 1990, p.1165–1175.

[3] J. Muhm, W. Miller, R. Fontana, D. Sanderson, and M. Uhlenhopp.: “Lung cancer detected during a screening program using four-month chest radiographs,” Radiology, 1983, vol. 148, p.609–615,.

[4] N. Hayabuchi, W. Russel, J. Murakami, and H. Nishitani.: “Screening for lung cancer in a fixed population by biennial chest radiography,” Radiology, 1983, vol. 148, p.369–373,.

[5] Van Ginneken, B. M. ter Haar Romeny, and M. Viergever.: “Computer-aided diagnosis in chest radiography: Asurvey,” IEEE Trans. Med. Imag., 2001, vol. 20, p.1228–1241.

[6] T. Kobayashi, X.-W. Xu, H. MacMahon, C. Metz, and K. Doi.: “Effect of a computer-aided diagnosis scheme on radiologists’s performance in detection of lung nodules on radiographs,” Radiology, 1996, vol. 199, p.843–848.

[7] J.H. Austin, N.L. Mueller, P.J. Friedman, et al., “Glossary of terms for CT of the lungs: recommendation of the Nomenclature Committee of the Fleischner Society”, Radiology 1996, vol. 200,p.327-331

[8] http://www.nlhep.org

[9] Aristofanes C. Cilva, Paulo Cezar, Marcello Gattas, “Diagnosis of Lung Nodule using Gini Coefficient and skeletonization in computerized Tomography images”, ACM symposium on Applied Computing March 2004.

[10] Ayman El-Baz, Aly A. Farag, Robert Falk, Renato La Rocca, “Detection,Visualization and identification of Lung Abnormalities in Chest Spiral CT Scan:Phase-I”, International Conference on Biomedical Engineering,Cairo, Egypt, 12-01-2002

[11] N. A. Memon, A. M. Mirza, S.A.M. Gilani, “Segmentation of Lungs from CT Scan Imges for Early Diagnosis of Lung Cancer”, Proceedings of World Academy of Science, Engineering and Technology, aug 2006 ,vol 14.

[12] Weiling Cai, Songcan Chen, Daoqiang Zhang, “Fast and Robust Fuzzy C-Means clustering algorithms incorporating local information for image segmentation”, Pattern Recognition, 2007.

[13] Alga singla and Rajesh Mahra, “Design and analysis of Fuzzy Clustering algorithm for data partitioning applications”, IJVSPA, Volume 1(2), pp.52-56, May 2006.

[14] N.R.Pal, and J.C.Bezdek, “A mixed c-means clustering model”, In IEEE Int.Conf.Fuzzy Systems, Spain, 1997, p.11-21.

[15] J.C.Bezdek, “Pattern Recognition with Fuzzy Objective Function Algorithms”, Plenum Press, 1981.

[16] Rezaul.K.Begg, Marimuthu Palaniswami and Brendan Owen (2005), Support Vector Machines for Automated Gait Classification”, IEEE Transactions on Biomedical Engineering, Vol.52, No.5.

[17] S.Sivakumar and C.Chandrasekar, “Lung Nodule Segmentation through Unsupervised Clustering Models”, Procedia Engineering , vol.38,p. 3064-3073.