Volume 3, No. 7, July 2012

Journal of Global Research in Computer Science

ISSN-2229-371X

RESEARCH ARTICAL

Available Online at www.jgrcs.info

AN EFFECTIVE CLASSIFICATION OF BENIGN AND MALIGNANT NODULES USING SUPPORT VECTOR MACHINE

M.Gomathi*[1], Dr.P.Thangaraj[2]

*[1]Associate Professor, Department of MCA, Velalar College of Engineering and Technology
Erode 638 012, TamilNadu, India
mdgomathi@gmail.com

[2]Professor& Head, Department of CSE, Bannari Amman Institute of Technology, Erode

Abstract: Support Vector Machine (SVM) is a machine learning technique that trains the system with the known data; it analyzes and identifies the patterns. SVM can be used for classifying the medical data because of its simplicity. Real time lung images are taken for the study. Lung images are segmented to retrieve the region of interest and these regions or nodules are used for classification. Proper threshold values are decided for each feature and classification rules are framed. Then these rules are passed to the SVM classifier. In this paper, classifications of benign and malignant nodules are done using different SVM kernels and their performance measures are compared.

Keywords: Lung Cancer, Features, Classification, Support Vector Machine, SVM Kernels.

INTRODUCTION

The most common cancer that occurs for men and women is lung cancer. The report submitted by the American Cancer Society in 2003 indicates that lung cancer is the cause for about 13% of all cancer diagnoses and 28% for all cancer deaths. The survival rate for lung cancer analyzed in 5 years is just 15 %. If the disease is identified while it is still localized, this rate increases to 49%. However, only 15% of diagnosed lung cancers are at this early stage.

Hence, it becomes necessary to detect the lung nodules in earlier stage (Gurcan, M.N, *et al.*, 2002) using chest Computer Tomography (CT) images. To achieve this, Computer Aided Diagnosis (CAD) system (Wiemker, R., *et al.*, 2003; Kanazawa, K., *et al.*, 1998) is very essential. Radiologists can miss up to 30% of lung nodules in chest radiographs due to the background anatomy of the lungs which can hide the nodules. Computer aided diagnosis system assists the radiologists by doing preprocessing of the images and recommending the most possible regions for nodules. Identification of lung regions progresses through methods for removing the background regions in lungs that comprise the blood vessels, ribs and the bronchi. The images resulted will provide good chest structure that create better regions for nodule and can be additional classified depends on the characteristics like size, contrast and shapes. The basic rule dependent classifications on such characteristics tend to generate a lot of false positives.

To overcome these problems, the author proposed a Computer Aided Diagnosing (CAD) (Wiemker, R., *et al.*, 2003) system for detection of lung nodules (Armato, S.G, *et al.*, 2001). This paper initially apply the different image processing techniques such as Bit-Plane Slicing, Erosion, Median Filter, Dilation, Outlining, Lung Border Extraction and Flood-Fill algorithms for extraction of lung region. Then for segmentation Modified Fuzzy Possibilistic C-Means algorithm (Mohamed Fadhel Saad et al., 2009; Gomathi *et al.*, 2010) is used and for learning and classification Support Vector Machine is used.

RELATED WORK

There are different computer aided diagnosis system exists for detection of lung cancer (Ginneken, B.V., 2001). Some of those techniques are discussed in this section.

Yamomoto *et al*., 1996 proposed image processing for computer-aided diagnosis of lung cancer by CT (LSCT). LSCT is the newly built mobile-type CT scanner mainly for the purpose of mass screening of lung cancer. In this new LSCT technique, one main complexity is the raise in the image information to around 30 slices per person from 1 X-ray film. In order to ignore this problem, the author attempts to reduce the image information considerably to be displayed for the doctor with the help of image processing techniques.

Yeny Yim *et al.*, 2005 stated about Hybrid lung segmentation in chest CT images for computer-aided diagnosis. The proposed system contains three phases. In the first phase, lungs and airways are separated by an inverse seeded region growing and linked component labeling. In the second phase, trachea and large airways are eliminated from the lungs by three-dimensional region growing. In the final phase, exact lung region borders are obtained by subtracting the outcome of the second phase from that of the first phase.

Penedo *et al.*, 1998 put forth a computer-aided diagnosis scheme that depends on two-level artificial neural network (ANN) architecture. The initial artificial neural network performs the identification of suspicious regions in a low-resolution image. The input provided to the second artificial neural network is the curvature peaks calculated for all pixels in each suspicious region. This is find out from the reality that small tumors possess and identifiable signature in curvature-peak feature space, where curvature is the local curvature of the image data when sighted as a relief map. The outcome of this network is threshold at a particular level of importance to provide a positive identification.

Kanazawa *et al.*, 1996 described Computer aided diagnosis system for lung cancer based on helical CT images. This technique will reduce the time complexity and increase the diagnosis confidence. This method consists of an analysis stage and a diagnosis stage. In the analysis stage, the lung and pulmonary blood vessel regions are extracted and examine the features of these regions with the help of image processing methods. In the diagnosis stage, diagnosis rules are determined according these features, and identify the tumor regions using these diagnosis rules.

Lin *et al.*, 2002 provided a neural fuzzy model to formulate the diagnosis rules for identifying the pulmonary nodules. Initially, series of image processing methods like thresholding, morphology closing, and labeling to segment the lung area and obtain the region of interest are used. Next, three main features such as circularity, size of area, and mean brightness are obtained from region of interest and the nodules are detected with diagnosis rules that are formed with the help of neural fuzzy model.

Armato *et al.*, 2001 developed a fully automated computerized technique for the identification of lung nodules in helical computed tomography scans of the thorax. This technique is based on two-dimensional and three-dimensional analyses of the image data obtained during diagnostic CT scans. Lung segmentation carried out on a section-by-section process to create a segmented lung volume within which further analysis is carried out. Multiple gray-level thresholds are supplied to the segmented lung volume for producing a series of thresholded lung volumes. An 18-point connectivity technique is implemented to detect contiguous three-dimensional structures within every thresholded lung volume, and those structures that satisfy a volume criterion are choosen as initial lung nodule candidates. Morphological and gray-level features are calculated for every nodule candidate. After a rule-based technique is used to highly decrease the number of nodule candidates that corresponds to nonnodules, the features of other candidates are combined through linear discriminant analysis.

SVMs are used for classification of breast cancer (Ireaneus Anna Rejani et al. 2009; Mu, T. et al., 2005) in which the performance were evaluated using the measure sensitivity.

METHODOLOGY

After the segmentation is performed on the lung region, the features can be obtained from it for determining the diagnosis rule for detecting the cancer nodules in the lung region perfectly. The features that are used to generate diagnosis rules (Sammouda *et al.*, 2005) are Area of the candidate region, maximum drawable circle (MDC) inside the candidate region, Average intensity value of the candidate region

After the necessary features are extracted, diagnosis rules can be applied to detect the occurrence of cancer nodule. The threshold value T1 is set for area of region. If the area of candidate region exceeds the threshold value, then it is eliminated for further consideration. The threshold T2 is defined for value of maximum draw-able circle (MDC). If the radius of the drawable circle for the candidate region is less than the threshold T2, then that is region is considered as non cancerous nodule and is eliminated for further consideration. Applying this rule has the effect of rejecting large number of vessels, which in general have a thin oblong, or line shape. The value T3 and T4 are set as threshold for the mean intensity value of candidate region. If the average intensity value of candidate region goes below minimum threshold or goes beyond maximum threshold, then that region is assumed as non cancerous region.

These rules can be passed to SVM classifier in order to detect the cancer nodules.

SUPPORT VECTOR MACHINE (SVM)

SVM introduced by Cortes is generally used for classification purpose. SVMs are efficient learning approaches for training classifiers based on several functions like polynomial functions, radial basis functions, neural networks etc. It is considered as a supervised learning approach that produces input-output mapping functions from a labeled training dataset. SVM has significant learning ability and hence is broadly applied in pattern recognition.

SVMs are universal approximators which depend on the statistical and optimizing theory. The SVM is particularly striking the biological analysis due to its capability to handle noise, large dataset and large input spaces.

The fundamental idea of SVM can be described as follows:

a. Initially, the inputs are formulated as feature vectors.
b. Then, by using the kernel function, these feature vectors are mapped into a feature space.
c. Finally, a division is computed in the feature space to separate the classes of training vectors.

A global hyper plane is sought by the SVM in order to separate both the classes of examples in training set and avoid over fitting. This phenomenon of SVM is more superior in comparison to other machine learning techniques which are based on artificial intelligence. The mapping of the input-output functions from a set of labeled training data set is generated by the supervised learning method called SVM. In a high dimensional feature space, SVM uses a hypothesis space of linear functions which are trained with a learning technique from optimization theory that employs a learning bias derived from statistical learning theory. In Support Vector machines, the classifier is created using a hyper-linear separating plane. It provides the ideal solution for problems which are not linearly separated in the input space. The original input space is non-linearly transformed into a high dimensional feature space, where an optimal separating hyper plane is found and the problem is solved. A maximal margin classifier with respect to the training data is obtained when the separating planes are optimal.

For binary classification SVM determines an Optimal Separating Hyperplane (OSH) which produces a maximum margin between two categories of data. To create an OSH, SVM maps data into a higher dimensional feature space and carries out this nonlinear mapping with the help of a kernel function. Then, SVM builds a linear OSH between two classes of data in the higher feature space. Data vectors that are closer to the OSH in the higher feature space are known as Support Vectors (SVs) and include all data necessary for classification.

A kernel function and the parameters should be selected for constructing the support vector machine classifier. Here, three kernel functions are used to construct SVM classifiers:

a. Linear kernel function
b. Polynomial kernel function
c. Radial basis function

The most used kernel function for SVM is Radial Basis Function (RBF) because of their localized and finite responses across the entire range of real x-axis. The classification accuracy of RBF kernel was high; also, the bias value and the error rate of RBF kernel were small when compared to other kernels.

Cross-validation:

Cross-validation is a method for analyzing how the results of a statistical analysis will generalize to an independent data set. It is used in situations, where the goal is prediction, and to estimate how accurately a predictive model will perform in practice. One round of cross-validation includes dividing or partitioning a sample of data into complementary subsets, performing the analysis on one subset (training set), and validating the analysis on the other subset (validation set or testing set). Multiple rounds of cross-validation are performed using different partitions to reduce variability, and the validation results are averaged over the rounds.

Confusion Matrix:

Confusion matrix was used to calculate the performance of the classifier. Figure 1 shows the confusion matrix. It is a specific table that helps to visualize the performance of a learning algorithm. Each column of the matrix represents the predicted class, and each row represents the actual class.

True positives	**False positives**
False negatives	**True negatives**

Figure 1 Confusion Matrix

The elements on the diagonal represent the correctly predicted elements and off the diagonal represent the misclassified elements.

EXPERIMENTAL RESULTS

The extracted features are passed to the SVM Classifier with three different kernels Linear, Polynomial and RBF. The performance of the classifier is measured based on the parameters sensitivity (Figure 2) and specificity (Figure 3). The aim of this paper is to classify or identify the malignant nodules (true positives) and Benign nodules (true negatives). The results are tabulated below:

Sensitivity = True Positives / (True Positives + False Negatives)

Kernel	**Sensitivity (%)**			
	Training	**Standard Deviation**	**Testing**	**Standard Deviation**
Linear (Dot)	85.36	0.63	82.11	0.71
Polynomial	88.61	0.57	84.55	0.64
Radial Basis Function (RBF)	91.05	0.45	86.17	0.52

Figure 2: Sensitivity of the SVM Kernels

Specificity = True Negative / (True Negatives + False Positives)

Kernel	**Specificity (%)**			
	Training	**Standard Deviation**	**Testing**	**Standard Deviation**
Linear (Dot)	84.18	0.68	81.92	0.75
Polynomial	86.44	0.61	83.61	0.67
Radial Basis Function (RBF)	89.83	0.51	84.74	0.58

Figure 3 : Specificity of the SVM Kernels

CONCLUSION

This paper provides a method for classification of benign and malignant nodules using Support Vector Machine. The features are extracted from the lung images. Then, diagnosis rules are framed for the extracted features. Finally with the obtained diagnosis rules, the classification is performed to detect the occurrence of cancer nodules. For the purpose of evaluation, the different real time chest computer tomography images are used. The experiment shows that the usage of SVM with RBF Kernel results in better accuracy of classification.

REFERENCES

[1] Yamomoto, S., Jiang, H., Matsumoto, M., Tateno, Y., Iinuma, T. and Matsumoto, T., “Image Processing for Computer-Aided Diagnosis of Lung Cancer by CT (LSCT)”, IEEE Workshop on Applications of Computer Vision, Pp. 236 – 241, 1996.

[2] Yeny Yim, Helen Hong and Yeong Gil Shin, “Hybrid Lung Segmentation in Chest CT Images for Computer-Aided Diagnosis”, International Workshop on Enterprise networking and Computing in Healthcare Industry, Pp. 378 – 383, 2005.

[3] Penedo, M.G., Carreira, M.J., Mosquera, A. and Cabello, D., “Computer-Aided Diagnosis: A Neural-Network-Based Approach to Lung Nodule Detection”, IEEE Transactions on Medical Imaging, Pp: 872 – 880, 1998.

[4] Kanazawa, K., Kubo, M. and Niki, N., “Computer Aided Diagnosis System for Lung Cancer Based on Helical CT Images”, International Conference on Pattern Recognition, Vol. 3, Pp. 381 - 385, 1996.

[5] Wiemker, R., Rogalla, P., Blaffert, T., Sifri, D., Hay, O., Srinivas, Y. and Truyen, R. “Computer-Aided Detection

(CAD) and Volumetry of Pulmonary Nodules on High-Resolution CT Data”, 2003.

[6] Lin, D. and Yan, C., “Lung Nodules Identification Rules Extraction with Neural Fuzzy Network”, IEEE Processing on Neural Information, Vol. 4, Pp. 2049- 2053, 2002.

[7] Armato, S.G, Giger, M.L. and MacMahon, H., “Automated Detection of Lung Nodules in CT Scans: Preliminary Results”, Medical Physics, Vol. 28, Pp. 1552–1561, 2001.

[8] Ginneken, B.V., Romeny, B.M. and Viergever, M.A., “Computer-Aided Diagnosis in Chest Radiography: A Survey”, IEEE Transactions on Medical Imaging, Vol. 20, No. 12, 2001.

[9] Gurcan, M.N, Sahiner, B., Petrick, N., Chan, H., Kazerooni, E.A., Cascade, P.N. and Hadjiiski, L., “Lung Nodule Detection on Thoracic Computed Tomography Images: Preliminary Evaluation of a Computer-Aided Diagnosis System”, Medical Physics, Vol. 29, No. 11, Pp. 2552- 2558, 2002.

[10] Kanazawa, K., Kawata, Y., Niki, N., Satoh, H., Ohmatsu, H., Kakinuma, R., Kaneko, M., Moriyama, N. and Eguchi, K., “Computer-Aided Diagnosis for Pulmonary Nodules Based on Helical CT Images”, Compute. Med. Image Graph, Vol. 22, No. 2, Pp. 157-167, 1998.

[11] Yamamoto, S., Matsumoto, M., Tateno, Y., Linuma, T. and Matsumoto, T., “Quiot filter - A New Filter Based on Mathematical Morphology to Extract the Isolated Shadow, and its Application to Automatic Detection Of Lung Cancer in X-Ray CT”, IEEE Proceedings of ICPR ’96, 1996.

[12] Mohamed Fadhel Saad and Adel M. Alimi, “Modified Fuzzy Possibilistic C-Means”, Proceedings of the International Multi Conference of Engineers and Computer Scientists, Vol. 1, 2009.

[13] Sammouda, R., Jamal Abu Hassan, Sammouda, M. Abdulridha Al-Zuhairy and Hatem abou ElAbbas, “Computer Aided Diagnosis System for Early Detection of Lung Cancer Using Chest Computer Tomography Images” GVIP 05 Conference, 19-21 December 2005, CICC, Cairo, Egypt, pp. 1-8, 2005.

[14] Y. Ireaneus Anna Rejani, Dr. S.Thamarai Selvi, “Early Detection of Breast Cancer Using SVM Classifier Technique”, International Journal on Computer Science and Engineering Vol 1, Issue 3, pages 127-130, 2009.

[15] Mu, T.; Nandi, A.K.; “Detection of breast cancer using v-SVM and RBF networks with self organized selection of centers”, The 3rd IEE International Seminar on Medical Applications of Signal Processing, pages 47 – 52, 2005.

[16] M.Gomathi, P.Thangaraj “A Parameter based Modified Fuzzy Possibilistic C-Means Clustering Algorithm for Lung Image Segmentation”, Global Journal of Computer science and Technology, Vol.10 No.4, 2010.

[17] M.Gomathi, P.Thangaraj “A New Approach to Lung Image Segmentation using Fuzzy Possibilistic C-Means Algorithm”, International Journal of Computer Science and Information Security, Vol.7 No.3, 2010.