Volume 3, Issue 3, March 2013 ISSN: 2277 128X
International Journal of Advanced Research in Computer Science and Software Engineering
Research Paper
Available online at: www.ijarcsse.com

A Study of Detection of Lung Cancer Using Data Mining Classification Techniques

Ada
M.Tech CSE
Department of Computer Science and Engineering
Sri Guru Granth Sahib World University
Fatehgarh Sahib, Punjab, India.

Rajneet Kaur
Assistant Professor
Department of Computer Science and Engineering
Sri Guru Granth Sahib World University
Fatehgarh Sahib, Punjab, India.

Abstract--Lung Cancer is a disease of uncontrolled cell growth in tissues of the lung. Detection of Lung Cancer in its early stage is the key of its cure. In general, a measure for early stage lung cancer diagnosis mainly includes those utilizing X-ray chest films, CT, MRI etc. In many parts of the world widespread screening by CT or MRI is not yet practical, so that chest radiology remains in initial and most common procedure. Firstly, we will use some techniques are essential to the task of medical image mining, Lung Field Segmentation, Data Processing, Feature Extraction, Classification using neural network and SVMs. The methods used in this paper work states to classify digital X-ray chest films into two categories: normal and abnormal. Different learning experiments were performed on two different data sets, created by means of feature selection and SVMs trained with different parameters; the results are compared and reported.

Keywords--Data Mining, Lung Cancer, Classification, Neural Networks, Support vector machine.

I. INTRODUCTION

Lung Cancer is a major cause of Mortality in the western world as demonstrated by the striking statistical numbers published every year by the American Lung Cancer Society. They indicate that the 5-year survival rate for patients with lung cancer can be improved from an average of 14% up to 49% if the disease is diagnosed and treated at its early stage.
Medical images as an essential part of medical diagnosis and treatment were concentrating on these images for good. These images include prosperity of unseen information that exploited by physicians in making reasoned decisions about a patient. However, extracting this relevant hidden information is a critical first step to their use. This reason motivates to use data mining techniques capabilities for efficient knowledge extraction & find hidden lung parts [1].
Mining Medical images involves many processes. Medical Data Mining is a promising area of computational intelligence applied to an automatically analyze patients records aiming at the discovery of new knowledge useful for medical decision-making. Induced knowledge is anticipated not only to increase accurate diagnosis and successful disease treatment, but also to enhance safety by reducing errors. The methods in this paper classify the digital X-ray chest films in two categories: normal and abnormal. The normal ones are those characterizing a healthy patient. The abnormal ones include Type of lung cancer; we will use a common classification method namely SVMs & neural networks.

II. METHODS AND TECHNIQUES

A. Lung Field Segmentation

Segmentation methods include the hidden parts in the lung area and avoid assumptions regarding chest position, size and orientation. It works with images where the chest is not always located in the central part of the images may be tilted and may have structural abnormalities. This algorithm detects the most visible lung edges by means of the first derivatives of Gaussian Filters taken at 4 different orientations. The edges thus detected provide an initial outline of the lung borders.

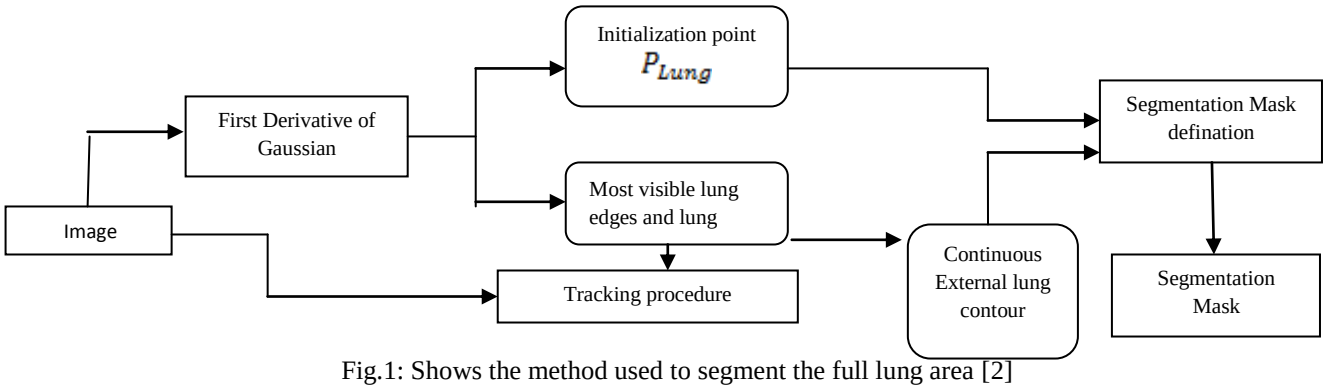

Fig.1: Shows the method used to segment the full lung area [2]

© 2013, IJARCSSE All Rights Reserved

P_{Lung} is the starting point for an edge-tracking procedure that works on 3 images representing the chest at 3 different levels of detail. These methods produce Segmentation Mask where hidden lung areas are excluded. Once the segmentation mask has been defined, a further method has been developed to find the separation between the hidden and the visible lung areas [2].

B. Data Preprocessing

Preprocessing phase of the images is necessary to improve the quality of the images and make the feature extraction phase more reliable. This phase consists of some processes. These processes contain Data Normalization, Data Preparation, Data Transformation, Data Cleaning and Data Formatting. Normalization techniques are necessary to combine the different image formats to a regular format. Data Preparation modifies images to present them in an appropriate format for transformation techniques. The image will transformed in order to obtain a compressed presentation. Segmentation completed to recognize regions of interest (ROI) for the mining task usually achieved using classifier systems [3].

C. Feature Extraction

Images usually have a huge number of features. It is important to recognize and extract interesting features for an exacting task in order to decrease the complexity of processing. Not all the attributes of an image are useful for knowledge extraction. Image processing algorithm used, which automatically extract image attributes such as local color, global color, texture, structure. The extraction of the features from an image can finished using a variety of image processing techniques. Based on this, the image is processed to look for a measurement that helps in selecting the pixels that correspond to the centers of the nodule. We localize the extraction process to very small regions in order to ensure that we capture all areas [3].

D. Classification

In recent years, many advanced classification approaches, such as neural networks, fuzzy-sets, expert system and SVM have been widely applied for image classification. In most cases, image classification approaches grouped as supervised & unsupervised machine learning approaches or parametric and non-parametric or hard and soft classification. The most used non-parametric classification approaches are neural networks, support vector machines & expert systems. Parametric classifier are robustness and easy to access for any image-processing software [1].

i. Neural Networks

An artificial neural network is a mathematical model based on biological neural networks. It consists of an interconnected group of artificial neurons and processes information using a connectionist approach to computation. Neurons are organized into layers. The input layer consists simply of the original data, while the output layer nodes represent the classes. Then, there may be several hidden layers. A key feature of neural networks is an iterative learning process in which data samples are presented to the network one at a time, and the weights are adjusted in order to predict the correct class label. Advantages of neural networks include their high tolerance to noisy data, as well as their ability to classify patterns on which they have not been trained [4]. A review of advantages and disadvantages of neural networks in the context of microarray analysis is presented [6].

The architecture of the neural network consists of three layers such as input layer, hidden layer and output layer. The nodes in the input layer linked with a number of nodes in the hidden layer. Each input node joined to each node in the hidden layer. The nodes in the hidden layer may connect to nodes in another hidden layer, or to an output layer. The output layer consists of one or more response variables [1].

A main concern of the training phase is to focus on the interior weights of the neural network which adjusted according to the transactions used in the learning process. This concept drives us to modify the interior weights while trained neural network used to classify new images.

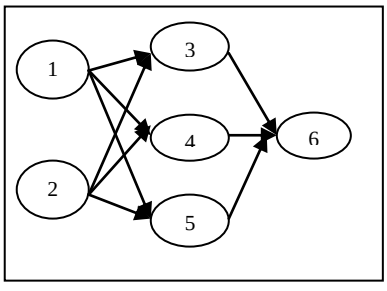

Fig.2 : A neural network with one hidden layer [1]

ii. Support Vector Machine

SVMs are based on the Structural Risk Minimization (SRM) principle from statistical learning theory. In their basic form, SVMs attempt to perform classification by constructing hyperplanes in a multidimensional space that separates the cases of different class labels. It supports both classification and regression tasks and can handle multiple continuous and nominal variables. Different types of kernels can be used in SVM models like linear, polynomial [5].

In the last years, SVMs have been widely investigated and used in a lot of different fields and for various classification tasks, due to their good performances. Learning algorithms such as neural network & SVMs, both trained with different parameters and input features, showed that SVMs produce the most robust results [7].

© 2013, IJARCSSE All Rights Reserved

Nevertheless, the results obtained were still not satisfactory due to high number of false positive candidates left after classification. With the aim of increasing classification performance, we calculated a set of 160 features including the 16 features which describe the shape and grey level of candidates [2].

Data Sets:-

For each region, we calculated the following five set of features:-

TABLEI DATA SETS

S.no	Data Items (Features)	Type of Data
1	19	shape and position of the region.
2	16	grey level distribution of the region pixels
3	6	radius value for candidate region
4	11	Compute at 11 different scales
5	108	Compute by Gaussian filters

Applying the feature selection technique, we selected 36 features to form the data set & its subset i.e. 18 features dataset. Since the number of true positives extracted (151) is much lower than the number of false positives (18916), both the 36 features and the 18 features Data Sets are very unbalanced; this explains the positive-enriched data sets that were built for training and testing the SVMs. For each data set, we separately considered the true positive and the false positive (negative) examples, and we randomly split the available positive data into 136 examples for training and 15 examples for testing, according to a train/test ratio of 9/1. From the set of negative data, we extracted without replacement a number of negative examples equal to 30 times the number of positive data obtaining, respectively, 136 × 30=4080 negative examples for the training set and 15 × 30=450 for the test data.

To test system performance when the number of positive examples used for training decreases, we ran other tests using a train/test ratio equal to 7/3, i.e., 106 examples for training and the remaining 45 for testing. In this case, we had 106 × 30=3180 in the training set & 45 × 30=1350 in the test set [2]. This process was repeated 10 times, obtaining 10 pairs of training and test sets.

Using the two data sets (the 36 features data set and the 18 features data sets), and the two ratios, 9/1 and 7/3, for splitting the positive examples into the train/test sets, we ran four experiments for each SVM [2].

III. CONCLUSION

In this paper, we are going to use some data mining classification techniques such as neural network & SVMs for detection and classification of Lung Cancer in X-ray chest films. Due to high number of false positives extracted, a set of 160 features was calculated and a feature extraction technique was applied to select the best feature. We classify the digital X-ray films in two categories: normal and abnormal. The normal or negative ones are those characterizing a healthy patient. Abnormal or positive ones include types of lung cancer. We will use some procedures also Data Preprocessing, Feature Extraction etc. In this paper we well use classification methods in order to classify problems aim to identify the characteristics that indicate the group to which each case belongs.

REFERENCES

[1] Zakaria Suliman Zubi and Rema Asheibani Saad, "Using Some Data Mining Techniques for Early Diagnosis of Lung Cancer," *Recent Researches in Artificial Intelligence, Knowledge Engineering and Data Bases,* Libya, 2007.

[2] Paola Campadelli, Elena Casiraghi, and Diana Artioli, "A Fully Automated Method for Lung Nodule Detection From Postero-Anterior Chest Radiographs," In *Proc. of IEEE TRANSACTIONS ON MEDICAL IMAGING*, VOL. 25, NO. 12, DECEMBER 2006.

[3] Jaba Sheela L and Dr.V.Shanthi, "An Approach for Discretization and Feature Selection Of Continuous-Valued Attributes in Medical Images for Classification Learning," *International Journal of Computer Theory and Engineering*, Vol. 1, No.2,June2009.

[4] V.Krishnaiah, Dr.G.Narsimha, Dr.N.Subhash Chandra. 2013, "Diagnosis of Lung Cancer Prediction System Using Data Mining Classification Techniques," *International Journal of Computer Science and Information Technologies*, Vol. 4 (1), 2013, 39 – 45.

© 2013, IJARCSSE All Rights Reserved

[5] Ankit Agrawal, Sanchit Misra, Ramanathan Narayanan, Lalith Polepeddi, Alok Choudhary, "A Lung Cancer Outcome Calculator Using Ensemble Data Mining on SEER Data," *BIOKDD 2011, August 2011, San Diego,* CA, USA, 2011.

[6] R.Linder, T.Richards and M. Wagner,"Microarray data classified by artificial neural networks," *METHODS IN MOLECULAR BIOLOGYCLIFTON THEN TOTOWA*-, 382:345, 2007.

[7] S. Xuejun, Q. Wei, and S. Dansheng, "Three-class classification in computer-aided diagnosis of breast cancer by support vector machine," *Proceedings SPIE Med. Imag.,* vol. 5370, pp. 999–1007, 2004

[8] S. S. Mohamed and M. M. A. Salama, "Computer-aided diagnosis for prostate cancer using support vector machine," *Proceedings SPIE Med. Imag.,* vol. 5744, pp. 898–906, 2005

[9] D. Glotsos, J. Tohka, P. Ravazoula, D. Cavouras, and G. Nikiforidis, "Automated diagnosis of brain tumours astrocytomas using probabilistic neural network clustering and support vector machines," *Int. J.Neural Syst*., vol. 15, pp. 1–11, 2005

[10] P. Campadelli, E. Casiraghi, and G. Valentini. "Lung nodules detection and classification," *In Proceedings. IEEE Int. Conf. Image Processing (ICIP2005),* pp. 1117-20, September, 2005.

© 2013, IJARCSSE All Rights Reserved