

Lung Cancer detection and Classification by using Machine Learning & Multinomial Bayesian

Mr. Sandeep A. Dwivedi¹, Mr. R. P. Borse¹, Mr. Anil M. Yametkar²

¹Department of E & TC, SAE, Kondhawa, Pune University.

²Department of E & TC, ARMEIT, Mumbai University.

Abstract: Image Processing is a technique to enhance raw images received from cameras/sensors placed on satellites, space probes and aircrafts or pictures taken in normal day-to-day life for various applications. Various techniques have been developed in Image processing during the last four to five decades. Most of the techniques are developed for enhancing images obtained from unmanned spacecrafts, space probes and military reconnaissance flights. Image Processing systems are becoming popular due to easy availability of powerful personnel computers, large size memory devices, graphics software etc.

Medical image segmentation & classification play an important role in medical research field. The patient CT lung images are classified into normal and abnormal category. Then, the abnormal images are subjected to segmentation to view the tumor portion. Classification depends on the features extracted from the images. We mainly are concentrating on feature extraction stage to yield better classification performance. Texture based features such as GLCM (Gray Level Co-occurrence Matrix) features play an important role in medical image analysis. Totally 12 different statistical features were extracted. To select the discriminative features among them we use sequential forward selection algorithm. Afterwards we prefer multinomial multivariate Bayesian for the classification stage. Classifier performance will be analysed further. The effectiveness of the modified weighted FCM algorithm in terms of computational rate is improved by modifying the cluster center and membership value updating criterion.

Objective of this paper is that

To achieve a perfect classification by multivariate multinomial Bayesian

Index Terms: Histogram Equalization, Image segmentation, feature extraction, neural network classifier, fuzzy c-means algorithm.

I. INTRODUCTION

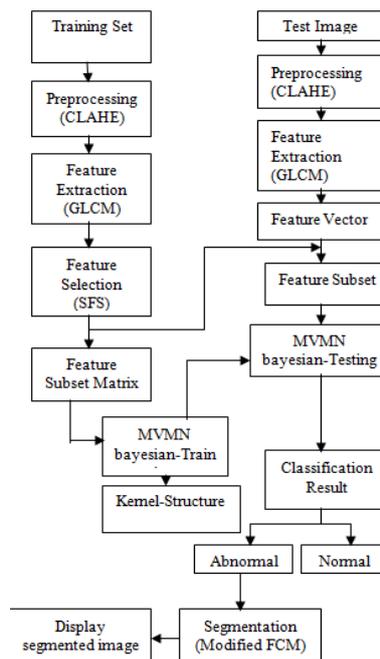
The early detection of lung cancer is a challenging problem, due to the structure of the cancer cells, where most of the cells are overlapped with each other. Classification is very important part of digital image analysis. It is a computational Procedure that sort images into groups according to their similarities. In this paper Histogram Equalization is used for preprocessing of the images and feature extraction process and neural network classifier to check the state of a patient in its early stage whether it is normal or abnormal. The manual analysis of the sputum samples is time consuming, inaccurate and requires intensive trained person to avoid diagnostic errors. The segmentation results will be used as a base for a Computer Aided Diagnosis (CAD) system for early detection of lung cancer which will improve the chances of survival for the patient. However, the extreme variation in the gray level and the relative contrast among the images make the segmentation results less accurate, thus we applied a thresholding technique as a pre-processing step in all images to extract the nuclei and cytoplasm regions, because most of the quantitative procedures are based on the nuclear feature.. Experimental analysis is made with dataset to evaluate the performance of the different classifiers. The performance is based on the correct and incorrect classification of the classifier. All experiments are conducted in WEKA data mining tool.

II. PROCEDURE OVERVIEW

A. Proposed Methodology

In our proposed method, we have to preprocess the given test image for reducing noise and to enhance the contrast. Afterwards, texture features (GLCM) will be extracted from it. In feature extraction stage, statistical measurements are calculated from the gray level co-occurrence matrix for different directions and distances. Among the various features extracted. We have to select the distinct features that will be utilized for classification purpose. For the selection of features SFS (Sequential Forward Selection) is used. Kernelised Bayesian is used to classify whether the test image comes under normal and abnormal.

B. System Architecture



C. System Requirements

1. Hardware Specification
 - Pentium IV – 2.7 GHz
 - 1GB DDR RAM
 - 250Gb Hard Disk
2. Software Specification
 - Operating system: Windows 7
 - Language : Matlab
 - Version : 7.9

III. PREPROCESSING

Contrast Enhancement:

- CLAHE- Contrast Limited Adaptive Histogram Equalization.
- CLAHE differs from ordinary adaptive histogram equalization (AHE) in its contrast limiting.
- The contrast limiting procedure has to be applied for each neighborhood from which a transformation function is derived.
- The contrast amplification in the vicinity of a given pixel value is given by the slope of the transformation function, which is Proportional to the slope of the CDF and therefore to the value of the histogram at that pixel value.
- CLAHE limits the amplification by clipping the histogram at a predefined value before computing the CDF.
- This limits the slope of the CDF and therefore of the transformation function. The value at which the histogram is clipped, the so-called clip limit, depends on the normalization of the histogram and thereby on the size of the neighborhood region.

While performing AHE if the region being processed has a relatively small intensity range then the noise in that region gets more enhanced. It can also cause some kind of artifacts to appear on those regions. To limit the appearance of such artifacts and noise, a modification of AHE called Contrast Limited AHE can be used. The amount of contrast enhancement for some intensity is directly proportional to the slope of the CDF function at that intensity level. Hence contrast enhancement can be limited by limiting the slope of the CDF. The slope of CDF at a bin location is determined by the height of the histogram for that bin. Therefore if we limit the height of the histogram to a certain level we can limit the slope of the CDF and hence the amount of contrast enhancement.

The only difference between regular AHE and CLAHE is that there is one extra step to clip the histogram before the computation of its CDF as the mapping function is performed. Hence CLAHE is implemented in the same function titled AHE in ahe.cpp. The program "AHE" takes an additional optional parameter which specifies the level at which to clip the histogram. By default no clipping is performed. Valid values for clipping fall in the range from 1 to 1/bins.

Following is the overview of the algorithm for this function:

1. Calculate a grid size based on the maximum dimension of the image. The minimum grid size is 32 pixels square.
2. If a window size is not specified chose the grid size as the default window size.
3. Identify grid points on the image, starting from top-left corner. Each grid point is separated by grid size pixels.
4. For each grid point calculate the histogram of the region around it, having area equal to window size and centered at the grid point.
5. If a clipping level is specified clip the histogram computed above to that level and then uses the new histogram to calculate the CDF.
6. After calculating the mappings for each grid point, repeat steps 6 to 8 for each pixel in the input image.
7. For each pixel find the four closest neighboring grid points that surround that pixel.
8. Using the intensity value of the pixel as an index, find its mapping at the four grid points based on their cdfs.
9. Interpolate among these values to get the mapping at the current pixel location. Map this intensity to the range [min:max) and put it in the output image.

IV. FEATURE EXTRACTION

A gray level co-occurrence matrix (GLCM) contains information about the positions of pixels having similar gray level values. A co-occurrence matrix is a two-dimensional array, P, in which both the rows and the columns represent a set of possible image values. A GLCM Pd[i,j] is defined by first specifying a displacement vector d=(dx,dy) and counting all pairs of pixels separated by d having gray levels i and j. The GLCM is defined by:

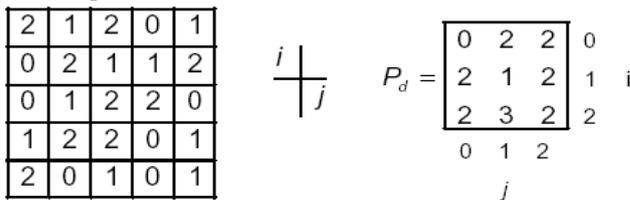
- where nij is the number of occurrences of the pixel values (i,j) lying at distance d in the image.
- The co-occurrence matrix Pd has dimension n x n, where n is the number of gray levels in the image.

From the co-occurrence matrix obtained, we have to extract the 21 different statistical features.

A gray level co-occurrence matrix (GLCM) contains information about the positions of pixels having similar gray level values.

- A co-occurrence matrix is a two-dimensional array, P, in which both the rows and the columns represent a set of possible image values.
- A GLCM Pd[i,j] is defined by first specifying a displacement vector d=(dx,dy) and counting all pairs of pixels separated by d having gray levels i and j.
- The GLCM is defined by:
 - where nij is the number of occurrences of the pixel values (i,j) lying at distance d in the image.
 - The co-occurrence matrix Pd has dimension n x n, where n is the number of gray levels in the image.

For example, if d=(1,1)



There are 16 pairs of pixels in the image which satisfy this spatial separation. Since there are only three gray levels, P [i,j] is a 3x3 matrix.

Algorithm:

- Count all pairs of pixels in which the first pixel has a value i, and its matching pair displaced from the first pixel by d has a value of j.
- This count is entered in the ith row and jth column of the matrix Pd[i,j]
- Note that P d [i,j] is not symmetric, since the number of pairs of pixels having gray levels [i,j] does not necessarily equal the number of pixel pairs having gray levels [j,i].

V. FEATURE SELECTION

Automatic feature selection is an optimization technique that, given a set of features, attempts to select a subset of size that leads to the maximization of some criterion function. Feature selection algorithms are important to recognition and classification systems because, if a feature space with a large dimension is used, the performance of

the classifier will decrease with respect to execution time and to recognition rate. The execution time increases with the number of features because of the measurement cost. The recognition rate can decrease because of redundant features and of the fact that small number of features can alleviate the course of dimensionality when the training samples set is limited, leading to overtraining. On the other hand, a reduction in the number of features may lead to a loss in the discrimination power and thereby lower the accuracy of the recognition system.

In order to determine the best feature subset for some criterion, some automatic feature selection algorithm can be applied to the complete feature space, varying the number of selected features from 1 to m.

VI. CLASSIFICATION BY MULTINOMIAL MULTIVARIATE BAYESIAN REFERENCES

A. Medical Image Classification

Classification Process

- 1) Training/Clustering Stage: the process of defining criteria by which patterns are recognized, developing a numerical description for each class
- 2) Classification Stage: each pixel in the image data set is categorized into the class it most closely resembles based on a mathematical decision rule
- 3) Output Stage: results are presented in a variety of forms (tables, graphics, etc.)

B. Supervised vs. Unsupervised Approaches

- Supervised - image analyst "supervises" the selection of image-regions that represent patterns/features that the analyst can recognize: Prior Decision
- Unsupervised - statistical "clustering" algorithms used to cluster the pixels, more
- Computer-automated: Posterior Decision.

C. Machine Learning and Classification

Machine Learning is a field of study that gives computers the ability to learn without being explicitly programmed. Given a training set, we feed it into a learning algorithm (like SVM, Artificial Neural Network, Logistic Regression, Linear Regression etc). The learning algorithm then outputs a function, which for historical reasons is called the hypothesis. Basically the hypothesis' job is to take a new input and give out an estimated output or class. The parameters that define the hypothesis are what are "learned" by using the training set. After the feature set has been computed for each pixel, it will be used by a classifier to decide whether each pixel represents a tumor pixel or a normal pixel. The classification stage has two components, a training phase and a testing phase. In the training phase, pixel features and their corresponding manual labels represent the input, and the output is a model that uses the features to predict the corresponding label. This training phase needs to be done only once, since the model can then be used to classify new data. The input to the testing phase is a learned model and pixel features without corresponding classes, and the output of the testing phase is the predicted classes for the pixels based on their features. A variety of classifiers including kNN, Decision Trees, Maximum Likelihood, Neural Networks, Ensemble Methods, Support Vector Machines, and Markov Random Fields. For most classifiers, assigning classes based on a model is computationally efficient, while initially learning the model can be computationally intensive. Sub sampling (using a subset of the full training data) is one method to ease the computational costs of the training phase, if the time needed to learn the model is prohibitively large. Although random sub-sampling can be used, spatial information can be used to produce a more strategic sub-sampling that will not degrade the quality of the learned model to the same extent as random sub-sampling. An obvious non-random sub-sampling strategy that uses spatial information is to sub-sample proportionally to the pixel's prior probabilities of being part of the brain mask, since few pixels outside the brain will be needed in training (assuming that a brain mask prior probability is used). Non random sub-sampling using spatial information could also be used to sub-sample normal areas that have large distances from tumor pixels, since these should exhibit fairly typical behavior and will likely not help significantly in learning a model that appropriately classifies ambiguous instances. "The support vector machine (SVM) is a universal constructive learning procedure based on the statistical learning theory developed by Vapnik, 1995. " Cherkassky and Mulier (1998). Classification with Support Vector Machines (SVM) has recently been explored by two groups for the task of brain tumor segmentation and represent a more appealing approach than ANN models for the task of binary classification since they have more robust (theoretical and empirical) generalization properties, achieve a globally optimal solution, and also allow the modeling of nonlinear dependencies in the features. As for MR brain image classification, it is typically solved by standard pattern recognition methods, with steps of feature extraction, feature dimensionality reduction, and feature based classification. Among these three steps, feature extraction is a crucial step. Once effective features have been extracted, feature dimensionality reduction and feature based classification can be straightforwardly completed by using suitable methods developed in machine learning area. For example, for feature dimensionality reduction, Principal Component Analysis (PCA) and feature selection techniques can be used, whereas for classification, the support vector machine (SVM) based

classifier or the linear discrimination analysis method can be applied. Aim of classification is to group items that have similar feature values into groups. Classifier achieves this by making a classification decision based on the value of the linear combination of the features.

D. Multinomial Bayesian Classification:

It is a generative (model based) approach, which offers a useful conceptual framework for Glaucoma images. Any kind of abnormalities can be classified, based on a probabilistic model specification. Features that describe data instances are conditionally independent given the classification hypothesis. Multivariate multinomial distribution for discrete data that fit assumes each individual feature follows a multinomial model within a class. The parameters for a feature include the probabilities of all possible values that the corresponding feature can take.

Bayes Rule is stated as follows,

$$P(h/d) = P(d/h) P(h) / P(d)$$

Understanding Baye's rule

d= data

h= hypothesis (model)

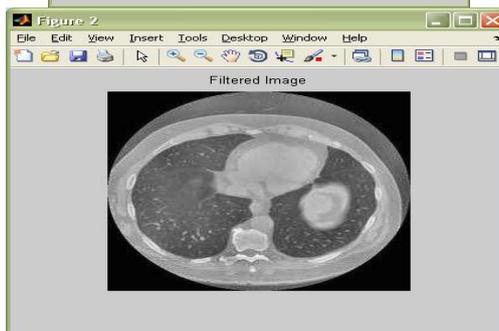
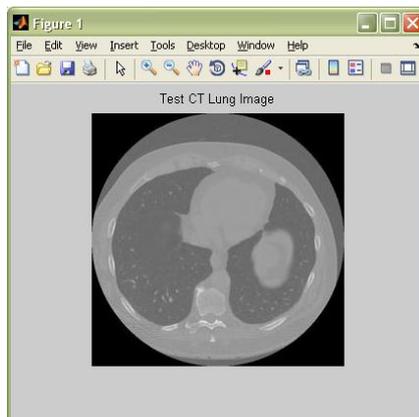
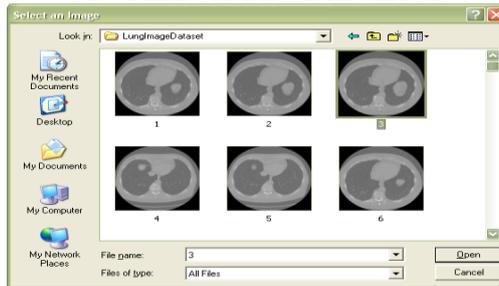
-rearranging

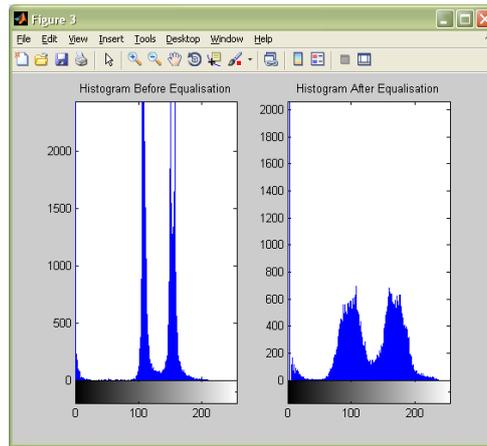
$$P(h/d) P(d) = P(d/h) P(h)$$

$$P(d, h) = P(d, h)$$

The same joint probability on both sides.

VI. PERFORMANCE ANALYSIS





Gray Level Distribution For Test Image

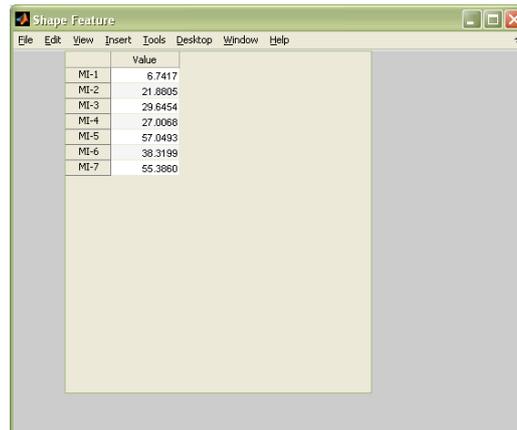
| Bin | Pixel count |
|-----|-------------|
| 1 | 0 |
| 2 | 1 |
| 3 | 2 |
| 4 | 3 |
| 5 | 4 |
| 6 | 5 |
| 7 | 6 |
| 8 | 7 |
| 9 | 8 |
| 10 | 9 |
| 11 | 10 |
| 12 | 11 |
| 13 | 12 |
| 14 | 13 |
| 15 | 14 |
| 16 | 15 |
| 17 | 16 |
| 18 | 17 |
| 19 | 18 |
| 20 | 19 |
| 21 | 20 |

Gray Level Distribution For Filtered Image

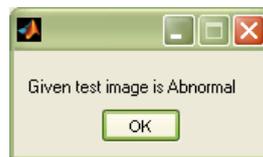
| Bin | Pixel count |
|-----|-------------|
| 1 | 0 |
| 2 | 1 |
| 3 | 2 |
| 4 | 3 |
| 5 | 4 |
| 6 | 5 |
| 7 | 6 |
| 8 | 7 |
| 9 | 8 |
| 10 | 9 |
| 11 | 10 |
| 12 | 11 |
| 13 | 12 |
| 14 | 13 |
| 15 | 14 |
| 16 | 15 |
| 17 | 16 |
| 18 | 17 |
| 19 | 18 |
| 20 | 19 |
| 21 | 20 |

Texture Features

| Feature | Value |
|--------------------|----------|
| Contrast | 0.4474 |
| Correlation | 0.9307 |
| Cluster Prominence | 324.9073 |
| Cluster Shade | -21.3916 |
| Dissimilarity | 0.3159 |
| Energy | 0.1244 |
| Entropy | 2.4412 |
| Homogeneity | 0.8599 |
| Homop | 0.8547 |
| Max. Prob | 0.1906 |
| Sosvh | 18.7853 |
| Autocorrelation | 18.6640 |



| | Value |
|------|---------|
| MI-1 | 6.7417 |
| MI-2 | 21.8805 |
| MI-3 | 29.6454 |
| MI-4 | 27.0068 |
| MI-5 | 57.0493 |
| MI-6 | 38.3199 |
| MI-7 | 55.3880 |



VII. FUTURE ENHANCEMENT

In future we aim to reduce the time taken for feature extraction of all dataset images. The training time will be reduced by sharing the work among different local processors. Parallel computing has been applied using multiple computational resources for feature extraction stage. Parallel computing also reduces the memory consumption by using a common shared memory. Work is shared among different workers. The number of workers that are going to share that work for a particular task is based on the number of samples. The single program multiple data construct allows seamless interleaving of serial and parallel programming. It lets us define a block of code to run simultaneously on multiple labs. This shared mechanism will definitely reduce execution time for feature extraction coding part.

VIII. CONCLUSION

The work in this research involves using kernelised Bayesian to classify the input which is CT lung image into normal and abnormal conditions. We intend to prove that this kernel technique will help to get more accurate result. Thus we have achieved high accuracy.

REFERENCES

- [1.] Guruprasad Bhat, Vidyadevi G Biradar , H Sarojadevi Nalini, “ Artificial Neural Network based Cancer Cell Classification (ANN – C3)”, Computer Engineering and Intelligent Systems, Vol 3, No.2, 2012.
- [2.] Almas Pathan, Bairu.K.saptalkar, “Detection and Classification of Lung Cancer Using Artificial Neural Network”, International Journal on Advanced Computer Engineering and Communication Technology Vol-1 Issue:1.
- [3.] Dr. S.A.PATIL, M. B. Kuchanur, ” Lung Cancer Classification Using Image Processing,” International Journal of Engineering and Innovative Technology (IJEIT) Volume 2, Issue 3, September 2012.
- [4.] Mokhled S. AL-TARAWNEH, “Lung Cancer Detection Using Image Processing Techniques,” Leonardo Electronic Journal of Practices and Technologies Issue 20, January-June 2012, p. 147-158.
- [5.] Fritz Albrechtsen, “Statistical Texture Measures Computed from Gray Level Cooccurrence Matrices,” International Journal of Computer Applications, November 5, 2008.
- [6.] Taranpreet Singh Ruprah, “Face Recognition Based on PCA Algorithm,” Special Issue of International Journal of Computer Science & Informatics (IJCSI), 2231–5292, Vol.- II, Issue-1, 2.
- [7.] ZAKARIA SULIMAN ZUBII, 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, 2011.
- [8.] Balaji Ganeshan, Sandra Abaleke, Rupert C.D. Young, Christopher R. Chatwin, Kenneth A. Miles, “Texture analysis of non-small cell lung cancer on unenhanced computed tomography: initial evidence for a relationship with tumour glucose metabolism and stage,” Cancer Imaging , v.10(1): 137–143, 2010 July 6.
- [9.] Lynne Eldridge MD. (2013, March 22). Lung Cancer Survival Rates by Type and Stage [Online]. Available:<http://lungcancer.about.com/od/whatislungcancer/a/lungcancersurvivalrates.htm>.
- [10.] Morphological Operators, CS/BIOEN 4640: Image Processing Basics, February 23, 2012.
- [11.] Image Processing – Laboratory 7: Morphological operations on binary images, Technical University of Cluj-Napoca, Computer Science Department.