Volume : 3 | Issue : 4 | May 2013 ISSN - 2250-1991

Research Paper Engineering

Minimal Feature Set Extraction for Classification of Lung Cancer CT-Scan Images

* Ada

* M.TECH, Student, Dept. of CSE, S.G.G.S.W.U., Fatehgarh Sahib(Punjab), India

ABSTRACT

This paper presents a comparison among the different classifiers Neural Network Classifier, Naive Bayes (NB), Support vector machine (SVM) on lung cancer data set by using classification parameters TP Rate (True Positive), FP Rate (False Positive), Mean Absolute Error and Root Mean Square Error to get the minimal feature set. All experiments are conducted in WEKA data mining tool.

Keywords : Classification, Lung Cancer, Lung Cancer dataset, Features.

INTRODUCTION

Lung cancer is considered to be the main cause of cancer death worldwide, and it is difficult to detect in its early stages because symptoms appear only in the advanced stages causing the mortality rate to be the highest among all other types of cancer. There are many techniques to diagnose lung cancer, such as Chest Radiography (x-ray), computed Tomography (CT), Magnetic Resonance Imaging (MRI scan) and Sputum Cytology [1]. Here, we used the CT-Scans of lung and choose the best features in images with highest priority to further classify the dataset.

Data mining and machine learning depend on classification which is the most essential and important task. Many experiments are performed on medical datasets using multiple classifiers and feature selection techniques. A good amount of research on lung cancer datasets is found in literature. Many of them show good classification accuracy [2]. Classifiers: In this experiment, the SMO from Support vector machine, Naïve Bayes from Bayesian Networks and Multilayer Perceptron (MLP) from Neural Networks are chosen to be compared to get the minimal feature set. These classifiers along with many other classification algorithms are implemented in Weka data mining tool.

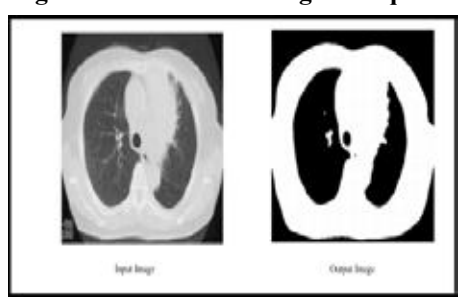
PRE-PROCESSING OF IMAGES

The image Pre-processing stage starts with image enhancement; the aim of image enhancement is to improve the interpretability or perception of information included in them image for human viewers, or to provide better input for other automated image processing techniques.

Image enhancement techniques can be divided into two broad categories: Spatial domain methods and frequency domain methods. Unfortunately, there is no general theory for determining what "good" image enhancement is when it comes to human perception. If it looks good, it is good. However, when image enhancement techniques are used as preprocessing tools for other image processing techniques, the quantitative measures can determine which techniques are most appropriate.

Lung Cancer Detection Using Image Processing Techniques processing tools for other image processing techniques, the quantitative measures can determine which techniques are most appropriate [3]. In the image pre-processing stage we used the Histogram Equalization.

Figure 1: Shows the Histogram Equalization on CT



scan image

FEATURE EXTRACTION

Now we extracted the 16 features of CT-scan images by using GLCM (grey level co-occurrence matrix) and binarization approach [4] [5] in MATLAB.

Features are:-

1. Contrast
2. Energy
3. Entropy
4. Homogeneity
5. Maximum Probability
6. Correlation
7. Cluster shade
8. Cluster Prominence
9. Dissimilarity
10. Autocorrelation
11. Sum variance
12. Sum Entropy
13. Difference Variance
14. Difference Entropy
15. Information Measure
16. Number of white pixels

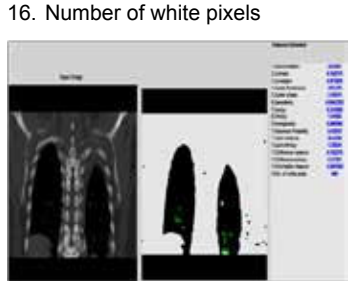


Figure 2: Calculated values of extracted features

DATASET
Now dataset is created of 350 images with values of 16 features extracted and a class attribute. It is in the form of .mat file.

CLASSIFICATION
Classification is a process which is used to categorize the data (XML, images, text etc) into different groups ("classes") according the similarities between them. Image classification is defined as the process to classify the pixels of images into different classes according to similarity [8].

The data is divided into different clusters and we saved the cluster assignment file in 'ARFF' format, whose last attributes shows the cluster assignment. The generated clustered file is used as input for classification in the next phase. Algorithms 'Neural Network', 'Naïve Bayes', 'SVM' has been implemented and results are recorded and studied for analysis purpose. In this the information gain of attributes is calculated and then put as input for the classification. Improved results have been obtained on our dataset.

IMPLEMENTATION
Lung Cancer CT-scans has been collected and after image processing total 16 features have been extracted from the images. A database file of 350 tuples and 16 attributes has been made in ASCII in CSV format, then conversion of this file to ARFF file is done. ARFF files are readable in Weka [9]. The generated ARFF file is opened in Weka and then different processes like data cleaning, data processing and data transformation are applied on to the input database file. These steps act as pre-processing steps for the classification of data. Along with this attribute removal is also done. Classification is implemented using 'Neural Network', 'SVM', 'Naive Bayes' [7] and results are recorded and studied for finding the minimal feature set for classification.

In order to find out minimal features set for Lung Cancer Images in case of supervised classification, first step is to find out the learning rate of different algorithms. To find out the learning rate of different algorithms, the training is started from 1 % percentage split and keeps on increasing till 99% percentage split. Results of different algorithms have been recorded and analysed and interpretation has been done according to the analyses.

MINIMAL FEATURE SET FOR CLASSIFICATION
Training of 50% is required in case of classification. To find out minimal feature set [6], start the evolution from 2 attributes and increase the number by 2 in every step till all 16 attributes has not covered.

It has been observed from the plotted values that at 12 attributes all algorithms give maximum value of parameters [7]. After 11 attributes algorithms gives constant value as they get stabilized.

RESULTS AND COMPARISON

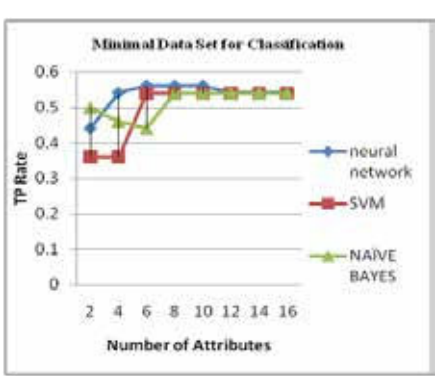


Figure 3: Minimal Feature Set for Classification using

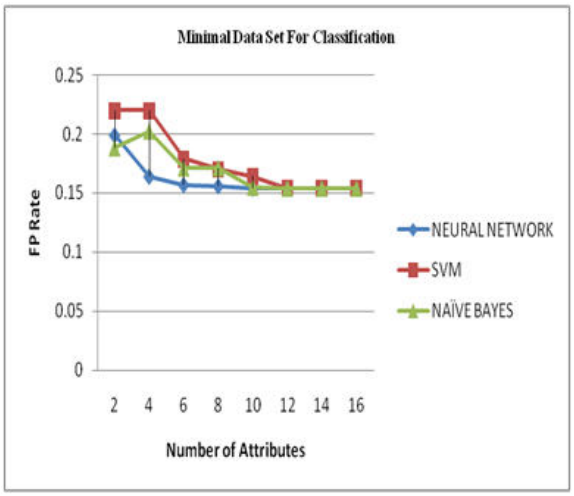


Figure 4: Minimal Feature Set for Classification using FP Rate

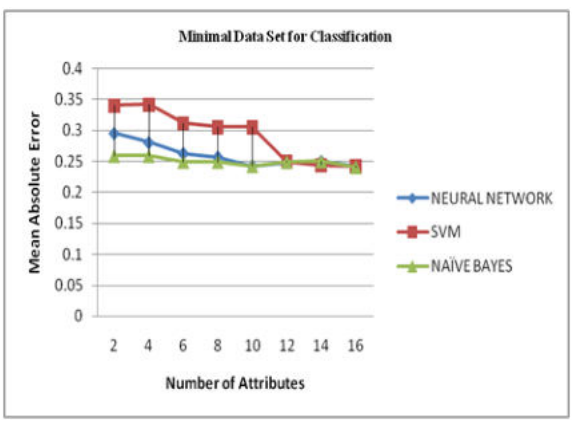


Figure 5: Minimal Feature Set for Classification using Mean Absolute Error

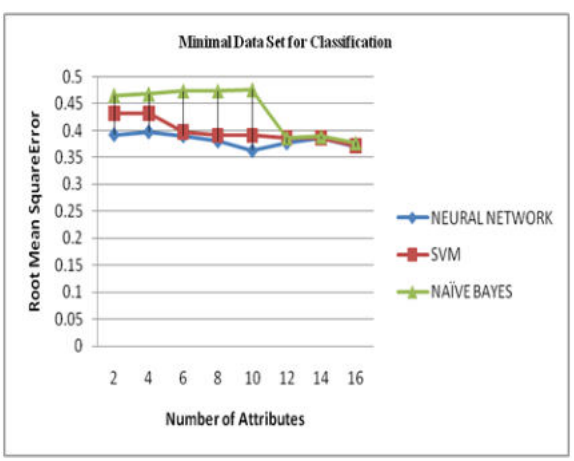


Figure 6: Minimal Feature Set for Classification using Root Mean Square Error

CONCLUSION
Images have been collected and then pre-processing performed on them. GLCM and Binarization approach is used for extraction of 16 features based on texture level in MATLAB. Then we made a dataset of these features with a class attribute in an ARFF format, so that we can load this data set in WEKA. Then, we apply the information measure to check the exact sequence of attributes. After that we start the evolution

from 2 attributes and increase the number by 2 in every step till all 16 attributes has not covered and apply different classification algorithms to get the exact minimal feature set on them. So it has been concluded that algorithms give the constant value after 12 features. It means that we can perform the classification on 12 attributes and can neglect the other 4 attributes on this lung cancer CT-Scan data set.

FUTURE WORK

The Experiments are performed only on 350 Lung Cancer CT-Scan images. Database can be extended and same methodology can be applied to the database containing images in thousands and many more. Only CT-Scan Lung Cancer data has been used, the same approach can be extended to different medical imaging technologies like X-ray, MRI etc. More features like Slice thickness etc can be calculated and feature set can be extended. Similarly for classification different combination of algorithms can be tried and results can be compared.

REFERENCES

[1] 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 . | [2] Gouda I. Salama, M.B.Abdelhalim, and Magdy Abd-elghany Zeid. (2012), "Breast Cancer Diagnosis on Three Different Datasets Using Multi-Classifiers," International Journal of Computer and Information Technology (2277 – 0764) Volume 01– Issue 01. | [3] Dr. S.A.PATIL, M. B. Kuchanur. (2012), " Lung Cancer Classification Using Image Processing," International Journal of Engineering and Innovative Technology (IJEIT) Volume 2, Issue 3. | [4] Mokhled S. AL-TARAWNEH. (2012), "Lung Cancer Detection Using Image Processing Techniques," Leonardo Electronic Journal of Practices and Technologies Issue 20, p. 147-158. | [5] Fritz Albregtsen. (2008), "Statistical Texture Measures Computed from Gray Level Coocurrence Matrices," International Journal of Computer Applications. | [6] Rajneet Kaur and Dr. Naveen Aggarwal .(2011)," Minimal Feature set for Unsupervised Classification of Knee MR Images," IJCSI International Journal of Computer Science Issues, Vol. 8, Issue 6, No 3. | [7] Bendi Venkata Ramana, Prof. M.Surendra Prasad Babu, Prof. N. B. Venkateswarlu. (2011), "A Critical Study of Selected Classification Algorithms for Liver Disease Diagnosis." International Journal of Database Management Systems (IJDMS), Vol.3, No.2. | [8] Wu, X. Kumar, V. Ross Quinlan, J. Ghosh, J. Yang, Q. Motoda, H. McLachlan, G. J. Ng, A. Liu, B. Yu, P. S. (2008), "Top 10 algorithms in data mining", Knowledge and Information Systems, vol 14; number 1, pages 1-37. | [9] Holmes, G.; Donkin, A.; Witten, I.H. (1994), "WEKA: a machine learning workbench", Intelligent Information system, proceedings of second Australian and new Zealand conference, pages 357-361. |