Available online at www.sciencedirect.com

ScienceDirect

Procedia Materials Science 10 (2015) 433 – 440

Procedia Materials Science

www.elsevier.com/locate/procedia

2nd International Conference on Nanomaterials and Technologies (CNT 2014)

Detection of Cancer in Lung With K-NN Classification Using Genetic Algorithm

P . Bhuvaneswari [a]*, Dr. A. Brintha Therese [b]

[a] *RajaRajeswari College of Engg,Bangalore ,Research Scolar,VIT University, Chennai,.India*
[b] *VIT University,Chennai,India*

Abstract

This paper focuses on early stage lung cancer detection. Genetic K-Nearest Neighbour (GKNN) Algorithm is proposed for the detection which is a non parametric method.This optimization algorithm allows physicians to identify the nodules present in the CT lung images in the early stage hence the lung cancer. Since the manual interpretation of the lung cancer CT images are time consuming and very critical, to overcome this difficulty the Genetic Algorithm method is combined with K-Nearest Neighbour (K-NN) algorithm which would classify the cancer images quickly and effectively .The MATLAB image processing toolbox based implementation is done on the CT lung images and the classifications of these images are carried out.The performance measures like the classification rate and the false positive rates are analyzed. In traditional K-NN algorithm, initially the distance between all the test and training samples are calculated and K-neighbours with greater distances are taken for classification. In this proposed method, by using Genetic Algorithm, K (50-100) numbers of samples are chosen for each iteration and the classification accuracy of 90% is achieved as fitness. The highest accuracy is recorded each time.

Peer-review under responsibility of the International Conference on Nanomaterials and Technologies (CNT 2014)

Keywords: Genetic Algorithm; Gabor filter;K-Nearest Neighbour;

1. Introduction

In nature, lung disease plays a major role in health issue. In any form of lung disease mainly the breathing gets affected, here are some common forms of lung diseases. Acute bronchitis, asthma, Chronic Obstructive Pulmonary

* Corresponding author. Tel.: 09448394177; fax: +91 80 2843 7373.
*E-mail address:*bhuvanasamuel@gmail.com

Peer-review under responsibility of the International Conference on Nanomaterials and Technologies (CNT 2014)
doi:10.1016/j.mspro.2015.06.077

Disease (COPD), chronic bronchitis, Emphysema, Acute respiratory distress syndrome (ARDS) and Lung cancer. As per World cancer report 2014 lung cancer is the most common cause of cancer-related death in men and women, and was responsible for 1.56 million deaths annually, as of 2012. The major causes of the lung diseases are smoking, inhaling the drugs, smoke and allergic materials. The computed tomography (CT) images assists in detecting the extreme of the lung diseases. For the analysis of the proposed method CT image is sufficient also the visibility of soft tissue is better .There are several types of lung cancer, and these are divided into two main groups: Small cell lung cancer and non-small cell lung cancer which has three subtypes: Carcinoma, Adeno carcinoma and Squamous cell carcinomas[1].

Neural Ensemble based Detection (NED) is proposed by Zhi-Hua Zhou et al [2] to identify lung cancer images in which one pass incremental learning is performed by adaptive neural classifier with high speed and accuracy. Mokhled S. AL-TARAWNEH proposed a technique to detect the features of accurate images by comparing the pixels percentage and mask-labeling and the time factor was considered to discover the abnormality issues present in target images. Image qual ity assessment as well as enhancement stage were adopted on techniques like Gabor filter within Gaussian rules. An algorithm is proposed by K.A.G.Udeshani, R.G.N.Meegama, and T.G.I.Fernando [3] which uses two steps of image processing. In the first step the separation of lung and in the second step a neural network is trained based on two types of inputs pixel based and statistical feature based and the recognition rates are tabulated and compared. Kerry A. Seitz et.al demonstrates the effectiveness and efficiency of content based image retrieval which can be improved by genetic algorithm [4].

In this proposed work for the noise removal and contrast enhancement the images are pre-processed to obtain the accurate enhanced images. Gabor filter is used in Feature extraction. The output values of Gabor filter are given to K-NN (Kernel Nearest Neighborhood) which is optimized by GA (Genetic Algorithm)

This paper is organized as follows: Section 2 focuses the proposed methodology. Section 3 explains the K-NN classification, Section 4 deals with Genetic algorithm and Section 5 presents the implementation of GA with K-NN along with the performance analysis and Section 6 deals the performance factors for better optimization of the images.

2. Methodology

The proposed method involves three stages are shown in Fig.1. Initially the CT lung images are pre-processed and segmented. Next stage is Feature extraction which is done by Gabor filter the third stage is classification with K-NN, and optimization by Genetic Algorithm.

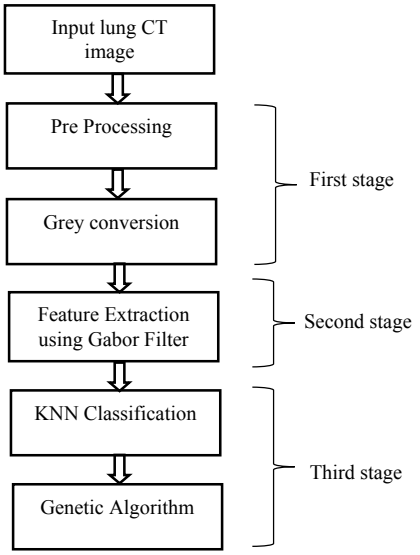

Fig . 1 Stages of Proposed Algorithm

In the proposed algorithm, enhancing the contrast of the input image through pre processing method is done by first converting the given input image to gray scale image. After enhancing the contrast of the image it is applied to Gabor filter to extract the feature contrast. The more usage of Gabor filter in image processing is texture analysis. Uncertainty Principle is used and provides precise time-frequency location. In both spatial and frequency domain these filters can be operated and their impulse response is defined by a sinusoidal wave multiplied by a Gaussian function. The convolution of the Fourier transform of the Gaussian function is the Fourier transform of a Gabor filter's response. The Gabor Filter's response contains real and imaginary component representing orthogonal directions shown by the below equations.

Complex

$$g\left(x,y;\lambda,\theta,\psi,\sigma,\gamma\right)=\exp\left(-\frac{x'^2+\gamma'^2y'^2}{2\sigma^2}\right)\exp\left(\iota\left(2\pi\frac{x'}{\lambda}+\psi\right)\right) \tag{1}$$

Real

$$g\left(x,y;\lambda,\theta,\psi,\sigma,\gamma\right)=\exp\left(-\frac{x'^2+\gamma'^2y'^2}{2\sigma^2}\right)\mathrm{Cos}\left(2\pi\frac{x'}{\lambda}+\psi\right) \tag{2}$$

Imaginary

$$g\left(x,y;\lambda,\theta,\psi,\sigma,\gamma\right)=\exp\left(-\frac{x'^2+\gamma'^2y'^2}{2\sigma^2}\right)\mathrm{Sin}\left(2\pi\frac{x'}{\lambda}+\psi\right) \tag{3}$$

Where

$$x'=xcos\theta+ysin\theta \tag{4}$$

$$y'=-xsin\theta+ycos\theta \tag{5}$$

The convolutions of the input image with the Gabor-filter kernels for all combinations of orientations and all phase-offsets with the input image are calculated by the Gabor filter which produces a 4D matrix contains image-coordinates, phase offset and orientation [5].The output of the Gabor filter is given to KNN classification which is optimized by Genetic Algorithm.

3. K-Nearest Neighbour Classification

In pattern recognition, the K-Nearest Neighbor algorithm (K-NN) is a non-parametric method used for classification and regression. In both cases, the input consists of the K closest training examples in the feature space. K-NN is a type of instance-based learning.

In K-NN Classification, the output is a class membership. Classification is done by a majority vote of neighbours. If K = 1, then the class is single nearest neighbour [6].

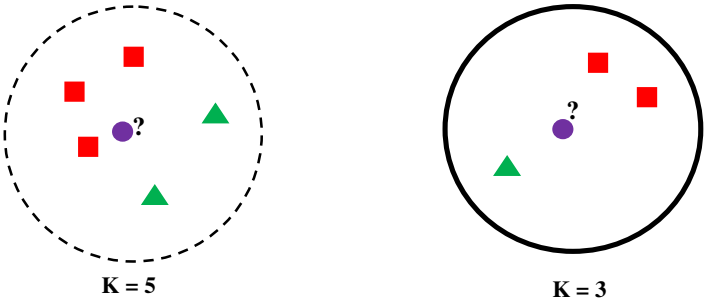

Fig .2. Classification of a test sample

A test sample classification is shown in the above Fig 2. Consider the test sample is a big dot located inside the circles which is classified either to the first class of triangles or to the second class of squares. If K=5 (dashed line circle) it is assigned to the second class because there are 3 squares and 2 triangles inside that circle. If K=3 (solid line circle) it is assigned to the second class because here 2 squares and 1 triangle inside that circle. It can be useful if the weight contributions of the neighbours are considered because the nearer neighbours contribute more than the distant ones. For example, in a common weighting scheme, individual neighbour is assigned to a weight of 1/d if d is the distance to the neighbour. The shortest distance between any two neighbours is always a straight line and the distance is known as Euclidean distance [7]. The limitation of the K-NN algorithm is it's sensitive to the local configuration of the data. The process of transforming the input data to a set of features is known as Feature extraction. In Feature space, extraction is taken place on raw data before applying K-NN algorithm. The Fig.3 narrates the steps involved in a K-NN algorithm.

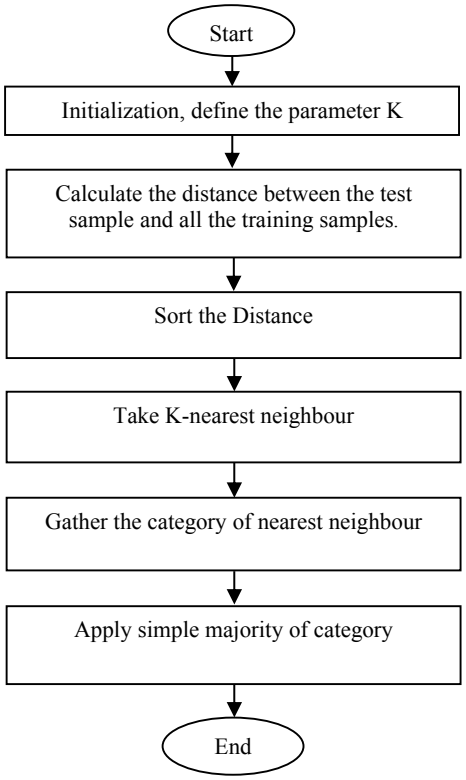

Fig . 3. K-NN Classification Algorithm

4. Genetic Algorithm

Genetic algorithms belong to the larger class of evolutionary algorithms, which generate solutions to optimization problems using techniques inspired by natural evolution, such as inheritance, mutation, selection, and crossover. [11]

4.1. Selection

In the nature, the selection of individuals is performed by survival of the fittest. If the individual is adapted to the environment more there is a bigger chance to survive and create an offspring and thus transfer its genes to the next population. In Evolutionary Algorithm the selection of the best individuals is based on an evaluation of fitness

function. Examples for such fitness function are the sum of the square error between the wanted system response and the real one; the distance of the poles of the closed-loop system to the desired poles, etc. If the optimization problem is a minimization one, then individuals with small value of the fitness function will have bigger chances for recombination and respectively for generating offspring.

4.2. Recombination

The first step in the reproduction process is the recombination (crossover). In it the genes of the parents are used to form an entirely new chromosome. The typical recombination for the GA is an operation requiring two parents, but schemes with more parents' are also possible. Two of the most widely used algorithms are Conventional (Scattered) Crossover and Blending (Intermediate) Crossover [9]

4.2.1. Conventional (Scattered) Crossover

In this recombination type, the parents exchange the corresponding genes to form a child. The crossover can be single- or multipoint as shown in Fig 4. (a) and (b). For recombination a bit *Mask* is used. The equation describing the process is:

$$C1 = Mask1 \,\&\, P1 + Mask2 \,\&\, P2 \quad (6)$$

$$C2 = Mask2 \,\&\, P1 + Mask1 \,\&\, P2 \quad (7)$$

Where
P1, P2 - Parent Chromosomes;
C1, C2 - Children Chromosomes (offspring individuals);
Mask1, Mask2 – bit Masks ;
Mask2 = NOT (Mask1)) & bit operation "AND"
For the example at Fig (4.b);

Mask 1 = [1 1 1 0 1 1 0 0 0]; Mask 2 = NOT (Mask1) = [0 0 0 1 0 0 1 1 1]; (8)
P1= [2 7 5 8 0 3 1 5 9]; P2= [8 8 4 5 1 6 9 7 1]; (9)

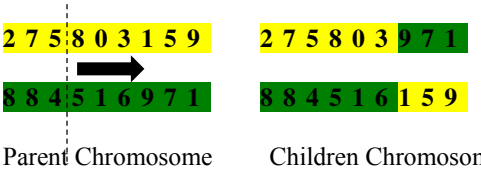

Fig . 4(a): Single Point Crossover

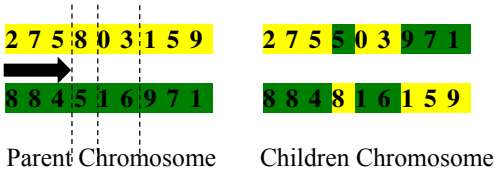

Fig . 4(b): Multi Point Crossover

4.3. Mutation

The offspring created by means of selection and crossover population can be further applied to mutation. In the terms of Genetic Algorithm, mutation means random change of the value of a gene in the population also, some elements of the DNA are changed. Those changes are caused mainly by mistakes during the copy process of the parent's genes. [10]

4.4. Implementation

4.4.1. String Representation

Here, encoding the chromosomes by real numbers; In each chromosome the number of genes represents the samples in the training set. Each gene will have k number of genes and 4 digits for vector index. For example, if K = 4, a sample chromosome may be written as follows:

0001 0102 0204 0302 0401 0500 0601 0702 0802 0901 1002 1101 1201

In the above string first two digits represents the attributes, the remaining two digits represents the value. Similarly the entire gene (4 digits) is encoded and now the initial population is generated hence we can apply genetic operators. With these K neighbours, the distance between each sample in the testing set is calculated and the accuracy is stored as the fitness values of this chromosome.

4.4.2. Selection

In the selection process each chromosome is selected from the mating pool as per the direction of the fittest concept of natural genetic systems. The common technique which implements the proportional selection strategy is Roulette wheel selection [12]. In proportional selection strategy, a chromosome is assigned a number of copies, which is proportional to its fitness in the population that go into the mating pool for further genetic operations is adopted.

4.4.3 Crossover

Crossover is a probability process which interchanges information between two parent chromosomes to generate two child chromosomes. Single point crossover with a fixed crossover probability of G is used in this paper. For chromosomes of length S, a random integer, called the crossover point, is generated in the range [1, S-1]. The chromosomes lying to the right portions of the crossover point are interchanged to produce two offspring.

4.4.3 Mutation

Each chromosome undergoes mutation with a fixed probability G. For binary representation of chromosomes, a bit position (or gene) is mutated by simply flipping its value. Since in this paper real numbers are considered, a random position is chosen in the chromosome and replace by a random number between 0-9. After the genetic operators are applied, the local maximum fitness value is calculated and compared with global maximum. If the local maximum is greater than the global maximum then the global maximum is assigned with the local maximum, and the next iteration is continued with the new population. The cluster points will be repositioned corresponding to the chromosome having global maximum. Otherwise, the next iteration is continued with the same old population. This process is repeated for N number of iterations. From the following section, it is shown that the below algorithm improves the cluster quality.

1. Choose K number of samples from the training set to generate initial population (P1).
2. Calculate the distance between training sets in each chromosome and testing samples, as fitness value.
3. Choose the chromosome with highest fitness value store it as Global maximum (Gmax), for iteration value 1to L
    - Perform reproduction
    - Apply the crossover operator.
    - Perform mutation and get the new population. (P2)
    - Calculate the local maximum (Lmax).
    - If Gmax < Lmax then
        Assign Gmax = Lmax;
        P1 = P2;
    - Repeat
4. Output – the chromosome which obtains Gmax has the optimum K-neighbours and the corresponding labels are the classification results. [8]

5. Results and Discussion

In this proposed method, we have successfully developed a solution for the detection of lung cancer nodules using image processing algorithms and neural networks. The algorithm is tested for five sets of cancer and non cancer lung CT images and shown in the below Fig.5 & Fig. 6. The input image and test image and their Gabor filter

outputs are displayed for both the cases. The K value, execution time and accuracy are calculated and tabulated. Refer Table . 1

Cancer Detected Images:

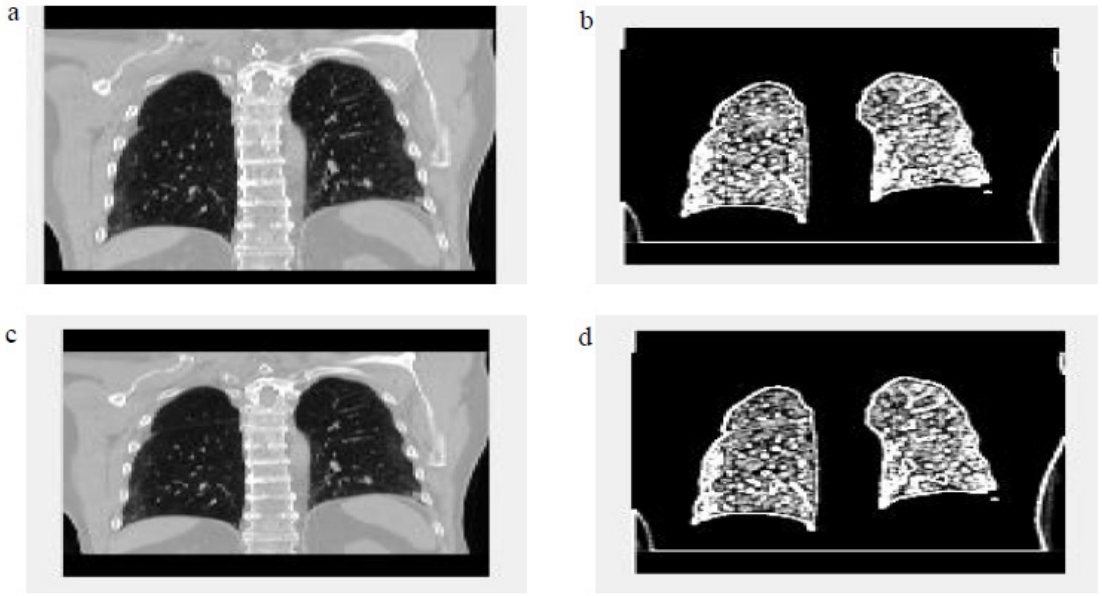

Fig .5. (a) Input image (b) Its Gabor Filtered Output (c) Test Image (d) Gabor Filter output of test image

Non Cancer Detected Images:

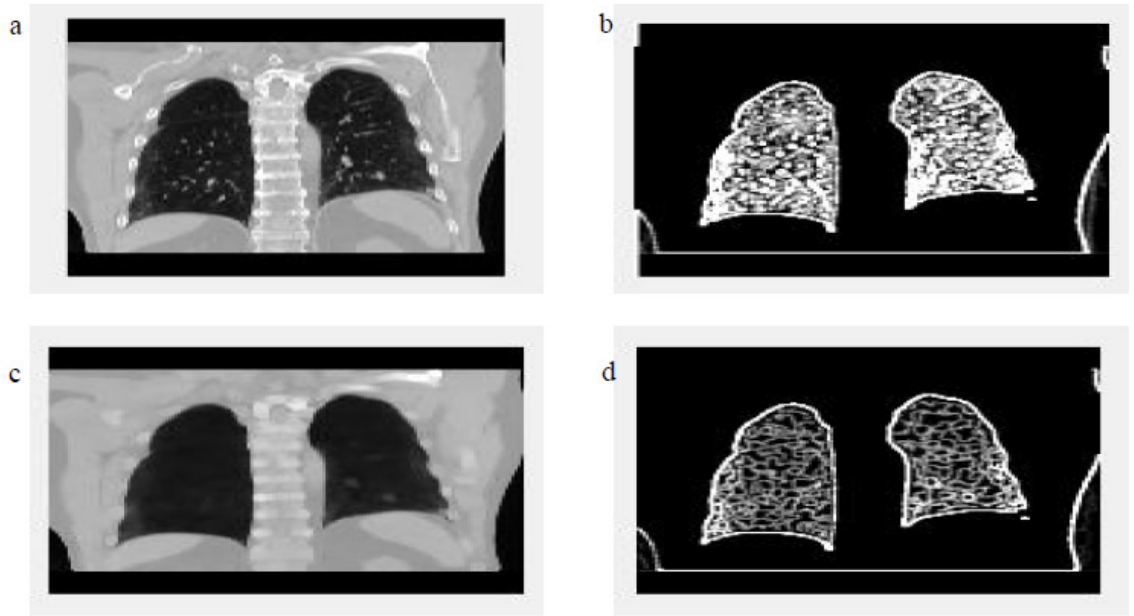

Fig . 6. (a) Input image (b) Its Gabor Filtered Output (c) Test Image (d) Gabor Filter output of test image

Table.1. Performance measures of K, Execution Time and Accuracy of Lung CT Images

Lung CT image	K	Execution Time(Sec)	Accuracy%
Cancer-1	52	3.65	89
Cancer-2	51	3.80	88
Cancer-3	53	3.60	88
Non Cancer-1	50	4.16	90
Non Cancer-2	49	4.25	89
Non Cancer-3	51	4.00	90

6. Conclusion

To combat the limitations of traditional K-NN, a novel method to improve the classification performance of K-NN using Genetic Algorithm (GA) is proposed in this paper. The proposed G-KNN classifier is applied for classification and similar k-neighbours are chosen at each iteration for classification by using GA, the test samples are classified with these neighbours and the accuracy is calculated for different number of K values to obtain high accuracy; hence the computation time of K-NN is reduced from the obtained results in this method.The MATLAB image processing toolbox based implementation is done on the CT lung images and the classifications of these images are carried out. The k value**,** execution time and accuracy is calculated and tabulated. Such early detection might be helpful for physicians.

Acknowledgements

Many thanks to Sri.A.C.Shanmugam, Chairman and Dr.M.S.BhagyaShekar, Principal of Rajarajeswari College of Engineering , Bangalore for giving an opportunity and encouragement to present this work . Also extend our gratitude to Colleagues and family members for their kind support.

References

Mokhled S. Al-Tarawneh 2012 Lung Cancer Detection Using Image Processing Techniques Leonardo Electronic Journal of Practices and Technologies Issue 20. pp. 147-158

Zhi-Hua Zhou, Yuan Jiang, Yu-Bin Yang, Shi-Fu Chen 2002 Lung Cancer Cell Identification Based on Artificial Neural Network Ensembles Artificial Intelligence Medicine Elsevier .Vol 24.Issue 1.p p 25-36

K.A.G.Udeshani,R.G.N.Meegama,T.G.I.Fernando 2011 Statistical Feature –based Neural Network Approach for the Detection of Lung Cancer in Chest X-Ray Images, International Journal of Image Processing.Vol 5 .Issue 4.pp 425-434

Kerry A. Seitz, Jr.a, Anne-Marie Giucab, Jacob Furstc, Daniela Raicuc . 2012. Learning Lung Nodule Similarity Using a Genetic Algorithm Proc. SPIE 8315,

W. Li , Kezhi Mao,Tianyou Chai .Selection of Gabor filter for improved texture feature Extraction. 2010 .17 th IEEE Conference on Image Processing, pp 361-364

Saravanan Thirumuruganathan, A Detailed Introduction to K-Nearest Neighbor (KNN) Algorithm. 2010. Wordpress.com

M. Akhil jabbara, B.L. Deekshatulub and Priti Chandrac,. 2013. Classification of Heart Disease Using K- Nearest Neighbor and Genetic Algorithm", Procedia Technology, Vol.10

N.Suguna,Dr.K.Thanushkodi 2010 An improved K-Nearest neighbour classification using genetic algorithm IJCSI Issue 4, Vol-7 pp 18-21

Shital Shah and Andrew Kusiak, 2007. Cancer gene search with data-mining and genetic algorithms", Computers in Biology and Medicine, Volume 37, Issue 2,

Temesguen Messaya, , Russell C. Hardiea, Steven K. Rogersb,. 2010 A new computationally efficient CAD system for pulmonary nodule detection in CT imagery, Medical Image Analysis,Vol 14,issue 3 pp 390–406.

Genetic Algorithm by David E. Goldberg, 2006. Pearson Education

J.Aruna Devi, Dr.V.Rajamani. 2011 An Evolutionary multi label classification using Associative Rule Mining for Spatial Preferences, IJCA AIT-Novel Approach and Practical Applications