Early Detection Of Lung Cancer Using Various Techniques

Neha Thombre
Student/ Department of Computer Engineering
Shree L R Tiwari College of Engineering, Mira Road,
Mumbai University, Mumbai, India
nehathombre@gmail.com

Seema Kolkur
Assistant Professor/ Department of Computer Engineering
Thadomal Sahani college of Engineering, Bandra
Mumbai University, Mumbai, India
seemakolkur@gmail.com

Abstract— In this paper, we are trying to compare two approaches which are used to identify cancer in early stages. First approach is about detection of cancer in lungs using KNN classification and then using GA and second approach is classification of heart disease using KNN and GA at the same time, where attributes are ranked based on GA and highest ranked attributes are selected and KNN is applied on that. Second approach gives more accurate results as compared to first one. Hence a modification is suggested in the first approach to rank the attributes by applying GA and higher ranked attributes should be used for KNN .This will save time and additional overhead and will give more accurate results.

Keywords—KNN,GA.

*****

I. INTRODUCTION

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 drugs, smoke, pollution and allergic materials. It is difficult to detect in its early stages because symptoms appear only at advanced stages causing the mortality rate to be the highest among all other types of cancer. There are four stages of lung cancer from I to IV with rising gravity. If the cancer is detected at stage I and it has no more 30 mm in diameter, then there is about 67% survival rate, and only less than 1% chance left for stage IV.It is advisable that early detection and treatment at stage 1 have high survival rate.

N. Suguna and Dr. K. Thanushkodi has proposed Genetic Algorithm (GA) is combined with k- Nearest Neighbor (KNN) algorithm called as Genetic KNN (GKNN), to overcome the limitations of traditional KNN[1].M.Akhil jabbar ,B.L Deekshatulua ,Priti Chandra approach combines KNN and genetic algorithm to improve the classification accuracy of heart disease data set. They used genetic search as a goodness measure to reduce redundant and irrelevant attributes, and to rank the attributes which contribute more towards classification. Least ranked attributes are removed[2].Sanwta Ram Dogiwal, Ȧ, Y.Singh Shishodi and Abhay Upadhyaya Ċ showed how Gabor filter can be used for image segmentation to extract features[4].

II. APPROACH USED 1

(Detection of Cancer in Lung with K-NN Classification Using Genetic Algorithm)

The manual interpretation of the lung cancer CT images are time consuming and very critical. In this proposed work they have tried to overcome this difficulty by using Genetic Algorithm method combined with K-Nearest Neighbor (K-NN) algorithm which would classify the cancer images quickly and effectively.

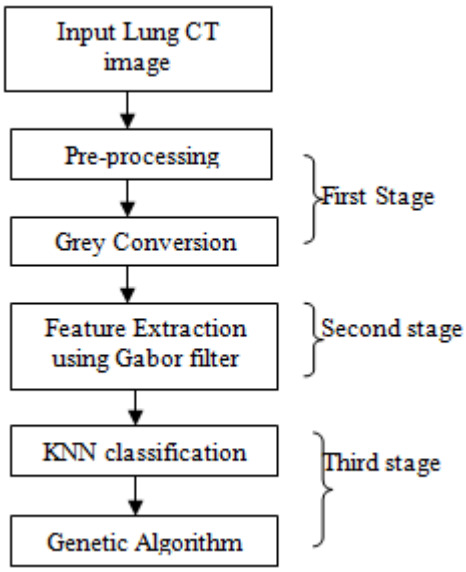

Figure 1. Methodology used

A. *GABOR FILTER*

In this proposed work for the noise removal and contrast enhancement the images are pre-processed to obtain the accurate enhanced images. Enhancing the contrast of the input image through preprocessing 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. Image enhancement divides into two major categories are (a) Spatial domain methods (b) frequency domain methods.

In spatial domain techniques- direct manipulation of image pixels. Frequency domain techniques- manipulation of Fourier transform or wavelet transform of an image. In image processing a Gabor filter is a linear filter used for edge detection. In the spatial domain, a 2D Gabor filter is a Gaussian kernel function modulated by a sinusoidal plane wave. The filter has a real and an imaginary component representing orthogonal directions.

The two components may be formed into a complex number or used individually.

Complex

$$g(x,y;\lambda,\theta,\psi,\sigma,\gamma)=\exp(-x'2+\gamma 2y'22\sigma 2)\exp(i(2\pi x'\lambda+\psi)) \quad (1)$$

Real

$$g \quad (x,y;\lambda,\theta,\psi,\sigma,\gamma)=\exp(-x'2+\gamma 2y'22\sigma 2)\cos(i(2\pi x'\lambda+\psi)) \quad (2)$$

Imaginary

$$g(x,y;\lambda,\theta,\psi,\sigma,\gamma)=\exp(-x'2+\gamma 2y'22\sigma 2)\sin(i(2\pi x'\lambda+\psi)) \quad (3)$$

x,y: coordinates specify the pixel value of an image at coordinates (x,y)

λ: represents the wavelength of the sinusoidal factor

θ: represents the orientation of the normal to the parallel stripes of a Gabor function

ψ:(PSI) is the phase offset (angular difference with respect to origin)

σ: is the sigma/standard deviation of the Gaussian envelope

γ: is the spatial aspect ratio, and specifies the ellipticity(deviation) of the support of the Gabor function

B. KNN Classification:

Nearest neighbour (KNN) is very simple, most popular, highly efficient and effective algorithm for pattern recognition and classification techniques introduced by Fix and Hodges which stores all cases and classify new cases based on similarity .KNN is a straight forward classifier, where samples are classified based on the class of their nearest neighbour. If the data set contains redundant and irrelevant attributes, classification may produce less accurate result. From training point to sample point distance is evaluated, and the point with lowest distance is called nearest neighbour.

The traditional KNN text classification has three limitations:

1. **High calculation complexity**: To find out the k nearest neighbour samples, all the similarities between the training samples must be calculated. When the number of training samples is less, the KNN classifier is no longer optimal, but if the training set contains a huge number of samples, the KNN classifier needs more time to calculate the similarities.

2**. Dependency on the training set**: The classifier is generated only with the training samples and it does not use any additional data. This makes the algorithm to depend on the training set excessively. It needs recalculation even if there is a small change in training set.

3**. No weight difference between samples**: All the training samples are treated equally there is no difference between the samples with small number of data and huge number of data. So it doesn't match the actual phenomenon where the samples have uneven distribution commonly.

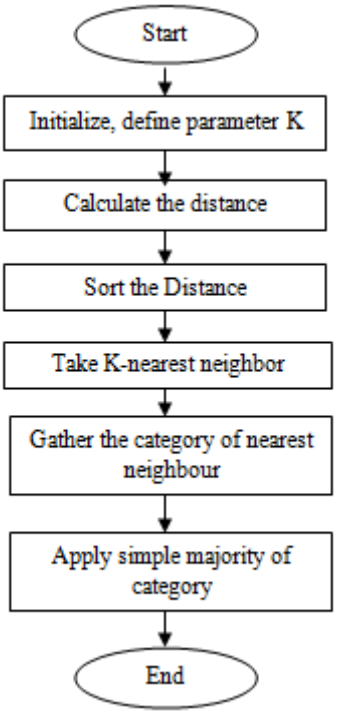

Figure 2. KNN classification

GENETIC ALGORITHM

Genetic algorithms (GA) was invented by **john Holland in** 1975.Genetic algorithms are useful for search and optimization problems.GA uses genetic as it's model as problem solving. Each solution in genetic algorithm is represented through chromosomes. Chromosomes are made up of genes, which are individual elements (alleles) that represent the problem. The collection of all chromosomes is called population. Generally there are three popular operators are use in GA.

Selection
Crossover
Mutation

Outline of the Basic Genetic Algorithm

1. **[Start]** Generate random population of *n* chromosomes.
2. **[Fitness]** Evaluate the fitness *f(x)* of each chromosome *x* in the population by decoding the chromosomes.
3. **[New population]** Create a new population by repeating following steps until the new population is complete.
    1. **[Selection]** Select two parent chromosomes from a population according to their fitness (the better fitness, the bigger chance to be selected).
    2. **[Crossover]** With a crossover probability cross over the parents to form new offspring

(children). If no crossover was performed, offspring is the exact copy of parents.

3. **[Mutation]** With a mutation probability mutate new offspring at each locus (position in chromosome).
4. **[Accepting]** Place new offspring in the new population.

4. **[Replace]** Use new generated population for a further run of the algorithm.
5. **[Test]** If the end condition is satisfied, **stop**, and return the best solution in current population.
6. **[Loop]** Go to step **2.**

To combat the limitations of traditional K-NN, a novel method to improve the classification performance of K-NN using Genetic Algorithm (GA) is suggested in this paper. The suggested 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 neighbors 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.

III. APPROACH USED -2

(Classification of Heart Disease Using K- Nearest Neighbor and Genetic Algorithm)

In this paper they have tried to classify heart disease data sets by 1) First dealing with evaluating attributes using genetic search 2) Part two deals with building classifier and measuring accuracy of the classifier.

Proposed algorithm

Step 1) load the data set

Step 2) Apply genetic search on the data set

Step 3) attributes are ranked based on their value

Step 4) selects the subset of higher ranked attributes

Step 5) Apply (KNN+GA) on the subset of attributes that maximizes classification accuracy

Step 6) calculates accuracy of the classifier, which measures the ability of the classifier to correctly classify unknown sample.

IV.COMPARISON OF BOTH TECHNIQUES:

Table I Comparison

steps	Lung Cancer	Heart disease detection
1	Load the IP lung CT IMAGE.	load the data set
2	Preprocessing of image	Apply genetic search on the data set
3	Gabor filter will extract features	attributes are ranked based on their value
4		select the subset of higher ranked attributes
5	Apply KNN to classify	Apply (KNN+GA) on the subset of attributes that maximizes classification accuracy
6	Apply **GA** to optimize	Calculate accuracy of the classifier, which measures the ability of the classifier to correctly classify unknown sample.

V. ANALYSIS

Table II Analysis of lung cancer using approach 1:

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

Sr no	Lung CT image	K	Execution Time (Sec)	Accuracy**%**
1	Cancer-1	52	3.65	89
2	Cancer-2	51	3.80	88
3	Cancer-3	53	3.60	88
4	Non Cancer-1	50	4.16	90
5	Non Cancer-2	49	4.25	89

Table III Analysis of heart disease using approach 2:

Sr no	Dataset Name	KNN+GA (our approach (k=1)
1	Weather data	100
2	Heart stalog	87.03
3	Lomography	100
4	Hypothyroid	95.75
5	Primary tumor	75.8
6	Heart disease A.P	95

VI. CONCLUSION

The traditional KNN text classification has three limitations:

1. High calculation complexity
2. Dependency on the training set
3. No weight difference between samples

In the first approach where we use Gabor filter to extract features. After extracting the features we are applying traditional way of KNN which has above mentioned limitation. Where as in second approach GA is used to rank

the features that eliminates less important features from consideration. Hence we recommend applying GA after the feature extraction to rank the features and avoid redundant and less important features.

Table IV SUGGESTED ALGORITHM

Steps	Suggestions in Lung Cancer Detection using K-NN Classification and Genetic Algorithm
1	Load the IP lung CT IMAGE
2	Preprocessing of image
3	Gabor filter will extract features
4	Apply GA to rank the features based on value
5	Apply KNN to classify
6	Apply GA to optimize

REFERENCE:

[1] N. Suguna1, and Dr. K. Thanushkodi:" An Improved k-Nearest Neighbour ClassificationUsing Genetic Algorithm ", IJCSI Issues, Vol. 7, Issue 4, No 2, July 2010

[2] M.Akhil jabbar B.L Deekshatulua Priti Chandra" Classification of Heart Disease Using K- Nearest Neighbour and Genetic Algorithm", Elsevier CIMTA 2013

[3] Sanwta Ram Dogiwal Ȧ, Y.Singh ShishodiaḂ and Abhay UpadhyayaĊ ," Study and Analysis of Different Gabor Filters for Image Enhancement and Segmentation",International Journal of Computer Systems, Vol. 01, Issue.01, September, 2014

[4] P. Bhuvaneswari a*, Dr. A. Brintha Therese "Detection of Cancer in Lung with K-NN Classification Using Genetic Algorithm" 2nd International Conference on Nanomaterials and Technologies (CNT 2014).

[5] "Lung Cancer Detection Using Image Processing Techniques",Mokhled S. Al-tarawneh, Leonardo Electronic Journal of Practices and Technologies, ISSN 1583-1078,Issue 20, January-June 2012.

[6] Khushbu Khamar **"**Short Text Classification Using KNN based on Distance Function", International Journal of Advanced Research in Computer and Communication Engineering Vol. 2, Issue 4, April 2013.

[7] R. Rajavignesh, Dr. G. Tholkappia Arasu, S. Vidivelli "Agent-Based Decision Support System using improved k-Nearest Neighbour Classifier for Effective Diagnosis of Heart Disease" Australian Journal of Basic and Applied Sciences ISSN:1991-8178.